APPRENTISSAGE SÉQUENTIEL : Bandits, Statistique et Renforcement.

Odalric-Ambrym Maillard

THÈSE DE DOCTORAT
de l'Université des Sciences et des Technologies de Lille

préparée au sein du LIFL, UMR 8022 Lille 1/CNRS
et du centre de recherche INRIA Lille - Nord Europe

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'USTL

Spécialité : **INFORMATIQUE**

APPRENTISSAGE SÉQUENTIEL :
Bandits, Statistique et Renforcement.

par
Odalric-Ambrym Maillard

Soutenue publiquement à **Villeneuve d'Ascq**, le **03 Octobre 2011** devant le jury composé de:

Philippe	Berthet	Université Toulouse III	Co-directeur
Olivier	Cappé	Telecom ParisTech	Président
Nicolò	Cesa-Bianchi	Università di Milano	Rapporteur
Pascal	Massart	Université de Paris-Sud	Examinateur
Rémi	Munos	INRIA Lille - Nord Europe	Directeur
Csaba	Szepesvári	University of Alberta	Rapporteur

Je dédie cette thèse à mon grand-père, René Jouannetaud, dont la force de vie n'eut d'égal que son infinie sagesse et sa bonté, ainsi qu'aux peuples mélanésiens du Vanuatu, dont la philosophie de vie exceptionnelle et la profonde humanité sont de vrais joyaux du monde.

Remerciements.

Les remerciements sont, je pense, un exercice primordial dans la vie d'un chercheur. Les gens que l'on rencontre nous influencent et nous aident de multiples façons, sans forcément que l'on s'en aperçoive, et c'est sans doute cette diversité de points de vue qui est le moteur le plus efficace pour imaginer de nouveaux concepts, trouver des solutions inédites, et garder un moral ainsi qu'une énergie constante.

J'aimerais remercier ici plusieurs personnes, en m'excusant d'avance très sincèrement pour les personnes que j'aurais pu oublier:

Mes directeurs de thèse. Mes premiers remerciements vont tout naturellement à Rémi[1], qui est devenu au cours de ces trois années, bien plus qu'un simple directeur de thèse, un véritable ami. Rémi, je suis toujours impressionné par ta manière d'être, chaleureuse et spontanée, par l'humilité et la recherche de simplicité dont tu fais preuve, par ton acharnement à comprendre et isoler chaque concept et par cet élan de travail et de bonne humeur que tu insuffles à l'équipe (les conversations à table sur les problèmes en cours, les questions "idiotes"...etc). Ce sont des choses qui me marqueront longtemps et que je véhiculerai avec moi. Je réalise la chance que j'ai eu de t'avoir pour directeur, et la liberté que tu m'as donné tout au long de cette thèse est sans aucun doute un cadeau précieux. Je ne perds d'ailleurs pas de vue l'idée que nous avions de travailler sur certains concepts de physique théorique, même si le cloisonnement scientifique - pourtant faible en France - ne nous a pas permis de nous lancer dans cette aventure trop risquée.

Philippe[2], c'est grâce à toi que j'ai choisi le chemin de la Statistique et que j'ai pu embrasser ce monde si passionnant et si vaste ! Je me souviens de tes cours denses et pourtant si limpides, dépassants parfois l'horaire de trois quart d'heure, se terminant dans le couloir; toute cette virtuosité et cette formidable énergie m'ont convaincu. Je sais aussi, sans l'avoir expérimenté aussi loin que toi cependant, combien la passion peut être dévorante - il y a tant de choses à faire et les journées sont trop courtes ! - mais je ne pense pas que dormir si peu chaque nuit soit bon sur le long terme. Je n'ai jamais vu quelqu'un d'autre animé de tant de ferveur, luttant ainsi de tout son être pour une cause noble, dans ton cas pour et par l'enseignement. C'est inestimable. Aussi je sais que tu ne m'écouteras pas. Durant cette thèse, la communication fut un peu compliquée, en particulier du fait de l'éloignement, peut-être aussi des conditions de recherche plus difficiles à Toulouse qu'à l'Inria, et de ton emploi du temps naturellement surchargé par les enseignements. Mais je reste confiant, car je sais que le fossé entre l'apprentissage par renforcement et la Statistique diminue de jour en jour et qu'il y a déjà moins de travail pour réunir les deux communautés qu'il y a trois ans. Cela dit il y a encore beaucoup à faire et déjà tellement de nouvelles questions passionnantes sont arrivées entre temps ! Une chose est certaine, je continuerai cet effort de rapprochement, ce grand écart délicat, et il se passera du temps avant que l'on ne s'ennuie !

[1] Rémi Munos

[2] Philippe Berthet

Les membres du jury. J'aimerais remercier sincèrement tous les membres du jury, Olivier Cappé, Nicolò Cesa-Bianchi, Pascal Massart et Csaba Szepesvári; je suis honoré que vous ayez tous répondu favorablement à mon invitation. Also I want to thank especially Nicolò Cesa-Bianchi and Csaba Szepesvári for accepting to review this Ph.D. dissertation (when they could go instead to the beach). Enfin, j'aimerais remercier tout spécialement Jean-Yves Audibert, qui n'a malheureusement pas pu se libérer pour le jour de la soutenance.

Mes co-auteurs extérieurs à l'équipe SequeL. La recherche serait bien plus ennuyeuse sans co-auteurs. Dans ma toute petite expérience de la recherche, j'ai déjà eu la chance de travailler avec des gens brillantissimes, et je suis heureux de voir que cela se poursuit en ce moment.

Nicolas[3], merci pour ta confiance, et ce, avant même le début de ma thèse. Gilles[4] ce fut un plaisir de travailler avec toi, et je recommencerai l'expérience sans hésiter (avec moins de stress sur la dernière ligne droite). Tu m'as apporté beaucoup, que ce soit en qualité rédactionnelle ou relationnelle, et je te remercie pour cela. Stéphane[5], après cette heureuse rencontre inattendue au détour d'un séminaire SMILE, nous commençons tout juste à travailler ensemble, et il y a de quoi faire ! Avec le travail de rédaction je n'ai pas encore pu m'investir comme je le voudrais, mais je promets de m'y mettre très sérieusement après. Aurélien[6], il semblait naturel que nous fassions cette fusion d'article avec Gilles. Je suis sûr que le résultat sera grand !

L'équipe SequeL. Now I would like to deeply thank Mohammad, who had to endure me during three years in the office 12, as well as Alessandro for all the nice discussions we had during this Ph.D. You are both brillant and accurate researchers and I am please to have worked and shared these years with you. Manuel, ce n'est pas si souvent que l'on rencontre une autre personne considérant que le monde est déterministe, et qu'en même temps cela ne doit affecter en aucun cas notre comportement, deux concepts considérés comme antagonistes par ceux qui ne les ont pas compris. D'une manière générale, ces petites discussions me manqueront sans doute. Merci à toi. Daniil, "the enlighted one", I am pleased that we finally found some time to work together, and I am positive that we will soon be able to grow a notion of Ryabko Decision Process.

J'aimerais également remercier toute l'équipe SequeL; en particulier Victor, Alexandra, ainsi que Phlippe, Jeremy, Sertán, Rémi, Jean-François sans oublier Sébastien qui est parti, Azalée et Mohammad qui nous ont rejoint récemment, et bien sûr Sandrine[7], que j'ai beaucoup sollicité pendant ces trois années, Géraldine et enfin Mélanie.

De futurs co-auteurs ? En dehors de l'équipe Sequel, j'ai été amené à rencontrer d'autres personnes qui m'ont marqué par leur enthousiasme scientifique et notamment Sylvain Ar-

[3]Nicolas Vayatis
[4]Gilles Stoltz
[5]Stéphane Gaiffas
[6]Aurélien Garivier
[7]Sandrine Catillon

lot, Alain Célisse, Jonas Kahn, Jens Kober, Guillaume Lecué, Adrien Saumard et Ohad Shamir que je salue ici très chaleureusement. Peut-être qu'un jour nous signerons un papier ensemble ?

Ceux dont on ne parle pas dans les articles. Avant de conclure, j'aimerais remercier ceux qui m'entourent et dont on ne parle pas dans les articles. Ma famille tout d'abord, ainsi que mes amis, qui font preuve d'une grande patience à mon égard et que j'avoue avoir quelque peu délaissé durant ces trois ans de thèse. Un immense merci à Thomas, le globe-trotteur infatiguable, pour m'avoir fait visiter la Croatie et l'Espagne, deux expériences inoubliables. Merci à toi, Bénédicte, pour tout cet amour. Et encore merci à tous les membres de mon Big-Band, le Jazz Orchestra Universitaire Lillois, pour toute cette magnifique expérience musicale ainsi qu'a tous ceux qui ont cru en ce projet un peu fou. Je passe la flambeau, confiant, car je suis convaincu que vous êtes prêt à vous passer de moi maintenant (en plus vous devez en avoir marre de m'avoir sur le dos) !

L'ENS et l'INRIA. Enfin je remercie très chaleureusement l'École Normale Supérieure de Cachan qui a financé cette thèse, ainsi que l'Institut National de Recherche en Informatique et en Automatique (INRIA) de Lille - Nord europe et le Laboratoire de Statistique et Probabilités (LSP) de l'Institut Mathématiques de Toulouse, qui m'ont accueilli au cours de ces trois années.

Pour finir, j'aimerais encore remercier ceux que j'ai oublié, et bien évidemment ceux qui liront ce manuscrit.

Résumé.

Cette thèse traite des domaines suivant en Apprentissage Automatique: la théorie des Bandits, l'Apprentissage statistique et l'Apprentissage par renforcement. Son fil rouge est l'étude de plusieurs notions d'adaptation, d'un point de vue non asymptotique : à un environnement ou à un adversaire dans la partie I, à la structure d'un signal dans la partie II, à la structure de récompenses ou à un modèle des états du monde dans la partie III.

Tout d'abord nous dérivons une analyse non asymptotique d'un algorithme de bandit à plusieurs bras utilisant la divergence de Kullback-Leibler. Celle-ci permet d'atteindre, dans le cas de distributions à support fini, la borne inférieure de performance asymptotique dépendante des distributions de probabilité connue pour ce problème. Puis, pour un bandit avec un adversaire possiblement adaptatif, nous introduisons des modèles dépendants de l'histoire et traduisant une possible faiblesse de l'adversaire et montrons comment en tirer parti pour concevoir des algorithmes adaptatifs à cette faiblesse.

Nous contribuons au problème de la régression en montrant l'utilité des projections aléatoires, à la fois sur le plan théorique et pratique, lorsque l'espace d'hypothèses considéré est de dimension grande, voire infinie. Nous utilisons également des opérateurs d'échantillonnage aléatoires dans le cadre de la reconstruction parcimonieuse lorsque la base est loin d'être orthogonale.

Enfin, nous combinons la partie I et II : pour fournir une analyse non-asymptotique d'algorithmes d'apprentissage par renforcement; puis, en amont du cadre des Processus Décisionnel de Markov, pour discuter du problème pratique du choix d'un bon modèle d'états.

Abstract.

This thesis studies the following topics in Machine Learning: Bandit theory, Statistical learning and Reinforcement learning. The common underlying thread is the non-asymptotic study of various notions of adaptation: to an environment or an opponent in part I about bandit theory, to the structure of a signal in part II about statistical theory, to the structure of states and rewards or to some state-model of the world in part III about reinforcement learning.

First we derive a non-asymptotic analysis of a Kullback-Leibler-based algorithm for the stochastic multi-armed bandit that enables to match, in the case of distributions with finite support, the asymptotic distribution-dependent lower bound known for this problem. Now for a multi-armed bandit with a possibly adaptive opponent, we introduce history-based models to catch some weakness of the opponent, and show how one can benefit from such models to design algorithms adaptive to this weakness.

Then we contribute to the regression setting and show how the use of random matrices can be beneficial both theoretically and numerically when the considered hypothesis space has a large, possibly infinite, dimension. We also use random matrices in the sparse recovery setting to build sensing operators that allow for recovery when the basis is far from being orthogonal.

Finally we combine part I and II to first provide a non-asymptotic analysis of reinforcement learning algorithms such as Bellman-residual minimization and a version of Least-squares temporal-difference that uses random projections and then, upstream of the Markov Decision Problem setting, discuss the practical problem of choosing a good model of states.

Foreword: To the layman reader.

One difficult exercise in research is to explain what we are actually doing to, say, "the guy in the street", i.e. someone who is not an expert of the field and maybe not even a scientist. In this introductory chapter, we try to explain and motivate what this thesis is about.

Mathematics, Computer Science, and "Informatics".

This thesis lies somewhere at the frontier between two very exciting domains. The first one is Mathematics, the second one is Informatics. Beyond the very naive separation between these two domains saying that Mathematics are interested in *theorems*[8] and *proofs* and that Informatics are interested in *computers*, *algorithms* and *complexity* (that is roughly speaking time and memory performance of algorithms), it is generally not so obvious to tell what is what, especially since these two first definitions are quite narrow.

Here, I intentionally use the word "Informatics" rather than the more common word "Computer science". The reason is that "Computer science" is a misleading word, as suggests the following quote attributed to Edsger Dijkstra: "*Computer science is no more about computers than astronomy is about telescopes.*" The french translation of Computer Science is "Informatique" and thus conveys a different meaning: that this is a science interested in information, or better said the information conveys by some objects, and not only in computers or algorithms. Moreover the word Informatics already exists, although being generally used in combination with other words, like in Bio-informatics.

More precisely, what I call "Informatics" here studies 1) how information is created or processed, 2) how information is transferred or altered between objects and 3) how to manage the objects of interest and retrieve information from them. For instance from a conventional Computer Science perspective, this is well handled by the abstract notion of a computer program that manages memory cells (bits) thanks to computer instructions written in some programming language and that produces a so-called trace - like a text, an image or the solution to an equation. Thus the study of programming languages and of *semantics*, a specific field of theoretical computer science, are clearly important in order to understand Informatics. But now the word Computer Science is not only misleading but also restrictive, as the previous example can be seen as the result of applying Informatics to some specific objects that are here memory cells (bits), while Informatics apply to more generic objects of interest and are thus much broader than what Computer Science suggests. Let us consider some random examples:

[8] In all this section, words in italics are technical words. There are not assumed to be known and their precise meaning should not prevent the reader from understanding the global message.

- For instance, let us consider that we apply Informatics to objects that are theorems. Then how we create information corresponds to the analysis of *axioms*, that are the basic statements assumed to be true and used as a starting point for reasoning. How information is transferred corresponds to the ways we combine theorems and make proofs: that is basic *logic* or *inference*. Now how we retrieve information from theorems is linked to deeper notions of logic that involve technical things like λ-*calculus* and *decidability*, with some famous difficulties pointed out by Godël in the 30's.

- For a more applied example, let us consider the result of applying Informatics to objects like proteins. This opens a very exciting field of research, directly linked with Biology. Indeed studying how proteins are created is one main question underlying *genomics* (before translation of *DNA*) and part of *proteomics* (after). Then the way they interacts with each other is studied by *proteomics*. Finally fields like e.g. Virology or Pharmacy study how one can manipulate them in order to build specific biological functions. More generally, applying Informatics to other "biological units" like neurons, cells or ecosystems, etc. results in the development of a new very active field of research called "Bioinformatics".

- For a last example, let us assume that the objects we consider are the rights of people, that is one important aspect of Law. Then one can use Informatics in order to study the creation of laws, the interaction between the rights by means of contracts and then the effects of the modifications of laws on the behavior of people. The study of such a complex dynamical system that consists of many interacting objects of different types - people, contracts, ownership, etc. - is definitely not easy.

What informatics bring Of course the benefit of Informatics here is the power of formalization, together with the development of powerful tools coming from *Graph* or *Domain theory* for instance, and the possibility to derive proofs, which is why the frontier between Mathematics is fuzzy. Actually it is even not important to tell what is what, if Informatics are a sub-field of Mathematics or if Mathematics are a sub-field of Informatics, the important thing is that Informatics enable us to analyze, understand and proof properties that concern a large diversity of topics, especially the not formalized one, and is thus a very helpful tool for the growth of precise knowledge.

The informatics of "learning" Now in this thesis, we are interested in the vague notion of *learning*. In order to apply Informatics to such a notion, we need some underlying object of interest. One way is to consider "data" or maybe sensors. Actually the underlying object of interest does not matter here since the notion of learning is itself a bit fuzzy. What is interesting is that with such objects, we roughly recover various aspects of the very broad field of research that is naturally called "*Machine learning*", and that is directly relevant to this thesis (the following words in italics refer to some key words in Machine Learning): For instance understanding how data is created or acquired is immediately identified as *sampling*

or *sensing*, and detecting structure in the data is the object of *clustering*, *coding theory*, or *graphical models*. The way data is altered and retrieved is addressed by problems like *regression* or *denoising*. Finally what happens when we manage the data, or act on sensors lies under the scope of what is called *active*, *sequential* and *reinforcement learning*.

Machine Learning, Artificial Intelligence and Statistical Theory.

As the last paragraph suggests, at a high level this work is interested in *learning* and more precisely in designing machines that can learn. For that reason, it is under the scope of Machine Learning, a field of research that is directly linked to the better known framework of *Artificial Intelligence*, and at the same time of *Statistical Theory*.

Artificial intelligence is challenging The difference with Artificial Intelligence, if there any, is that in Machine Learning, we want to design a machine that learns *something*, which means that the goal is given beforehand, and is fixed during all the learning process. On the other hand in Artificial Intelligence, we also would like the machine to be able to adapt on the fly to a change of the initial goal, thus to detect when we ask for a different goal, and to reuse, while continuing to learn, its past knowledge to target a new goal; this also means that we may want to measure how close is a new goal from previously identified goals. This is far more challenging than the questions classically addressed by Machine learning, and, naturally, a large amount of questions in this setting have not been answered so far. That said, this definition of Artificial Intelligence is only one amongst many (see Legg and Hutter (2007)), and some people may consider this distinction perhaps less relevant.

Statistical theory is exciting Machine Learning is also linked to Statistical Theory, for this field of research studies what can be deduced from observations. We will focus on designing and studying decision mechanisms for a machine, that we call decision *algorithms*. These algorithms form the "brain" of the machine and we need mathematical tools in order to design an analyze them carefully. More precisely, from a mathematical point of view, I here consider Machine Learning as being exactly *non-asymptotic* statistical theory, i.e. understanding what can be deduced when we are only allowed to get finitely many observations. Since there are many important things to understand from a non-asymptotic point of view, this second aspect of Machine Learning makes it a tremendously interesting field of research.

Reinforcement learning.

One important step towards addressing the challenge of Artificial Intelligence comes, to my mind, from the sub-field of Machine Learning called Reinforcement Learning, that enables to formalize many aspects of this challenge and to design algorithms that are theoretically sound. Reinforcement learning is based on three important notions: states, actions, and

rewards. The states and the rewards are determined by the environment, while the learning machine tries to choose the best actions in some sense.

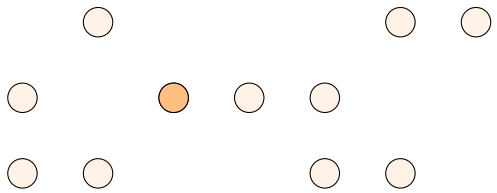


Figure 1: A set of states, and an example of a current state in orange.

States More precisely, in this setting, the relevant pieces of information about the machine at some time t are gathered into something called a *state*. A state is for instance the position of a robot in a room, or the configuration of the board at some instant at chess. There is generally a set of possible states (see Figure 1), and we can only be in one state at a time, that we call here the current state. For the sake of simplicity, the set of possible states - here the shape of the room, or the set of all possible board configurations - is generally assumed to be known, as well as the current state - thus we know our current position, or the current configuration. But in some practical applications, these two assumptions may be too strong, and one has to either deduce its own position if the set of states is known but not the current state, or even worse to infer a sensible set of states. The first situation typically appears when the robot is in a maze.

Actions and rewards The machine interacts with some environment by outputting *actions* and by receiving some *rewards* from the environment each time it plays an action; the reward measures the quality of the action. For the robot, the actions are typically go north/west/south/est, and a reward can be 3ml of oil if it moves towards a goal, 1ml if it moves almost towards it, an nothing else. An action for chess is for instance push pion D2 to D3, and a typical reward for an action can be: 1 if your action do mat, -1 if your opponent do mat right after your action, and 0 for all other situations.

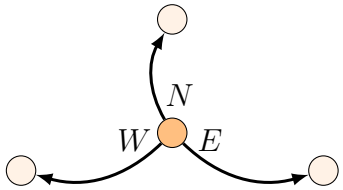


Figure 2: A current state, with 3 possible actions: W/N/E

The effect of playing one action is to move from the current state to another one. Note that a change of state can only appear when an action is played - this is actually a good

property for a state, and also that in general playing one action in some state may lead the machine to the very same state, in which case there is no move.

The environment The way a reward is given to a machine when it plays some action a in a state s is a property of the environment. This is the *structure of rewards*. Similary, the way we move from one state s to another one s' after playing some action a is also a property of the environment. We call it the *structure of transitions* between states. See figure 3 for an example of such structures.

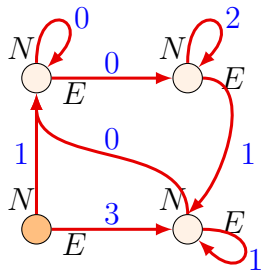


Figure 3: A possible structure of transition in red, and structure of rewards in blue.

Best actions Now, when the machine plays a sequence of actions, it moves at each time-step from one state to another and receives for each output action some reward. It thus receives a sequence of rewards that depends on the sequence of actions played as well as the initial state. We measure the quality of a sequence of actions by the sum of these rewards.

The goal of the machine is to learn, for each initial state, one right sequence of actions to choose in order to receive the highest sum of reward when playing the chosen sequence of actions. This is generally not easy especially when the structure of rewards and of transitions are not known in advance.

This Ph.D. dissertation is about the theory of Reinforcement Learning, with the goal to provide theoretically sound and numerically efficient decision-algorithms for that setting. Let us now give more practical applications for this important field of research.

Reinforcement learning: what for?

When discussing with people about your Ph.D., you are almost always asked, what is the real application of your thesis in the real-world? What is it for? Here I give some possible answers to this difficult and important question. When we talk about Reinforcement Learning, one naturally thinks about robots, which is indeed one direct application, although it also sounds like science fiction, fantasy, utopia etc. But actually, applications of Reinforcement Learning

go far beyond this scope, and I now present some of them that I consider to be important, maybe from the most obvious to the most surprising one.

Robot control on Mars. In this decision-making scenario, a planetary rover must explore a number of sites during a period of time without stopping to establish communication with Earth (Bernstein and Zilberstein (2001)). Using only information about its resource levels, and about the goals of the mission (rewards), the rover must decide which activities to perform and when to move from one site to the next (actions). Here the states gather information like the time remaining in the day, the current site and the successful activities performed in this site. Limited resources and non-deterministic action effects make the problem nontrivial.

Board game player. In this problem, the learning agent must learn how to play to some board game, like chess, backgammon or go. One example of action is for instance to move a pion at chess. The learning agent classically receives at the end of each game against an opponent a reward 1 if it wins, -1 it if loses, and 0 else. For instance, thanks to reinforcement learning, there exist now algorithms that can defeat all human beings at backgammon (Tesauro (1995)), achieve international level at chess (Baxter et al. (2001)) and some other games. The current challenge of the community is the game of Go. Indeed, for this game, if some algorithms do achieve national level on a small board 9×9, computers reach a level that is still way below the level of professional go-players on the full board 19×19. See Coulom (2007), Gelly and Silver (2007) for interesting research on this topic.

Elevator group control. In this setting, a system of elevators is controlled using a group of reinforcement learning algorithms, with one learning algorithm per elevator, and the goal is to minimize the waiting time of people when they call for a lifter. The team of elevators receives a global reward signal which appears noisy to each agent due to the effects of the actions of the other agents, the random nature of the arrivals and the incomplete observation of the state. This has been studied in Crites and Barto (1996) and tested in practice. Despite the difficulty of the task, the system was able to predict after few weeks of learning how to place elevators appropriately, anticipating some interesting features like where and when rush hours may occur.

Hydraulic Stock gestion. We consider a power supply system, like EDF for instance, based on thermo-hydraulic electricity production. In this problem, we need to produce precisely as much electricity as the consumer demand, using either hydraulic stocks, which is cheap but limited or thermic complexes, which is expensive but less limited in order to produce electricity. Our goal is to determine whether it is better to use the water stock now, or to keep it for a possible high consummation demand in the future. Thus, we need to manage the hydraulic stock in an efficient manner in order to maintain low costs while producing the amount of electricity required. See Aïd et al. (2003) for some application.

Web advertisement. Displaying advertisements on a web page is one direct application of the sub-field of reinforcement learning known as *Bandit games* named after slot machines in casinos, see chapter 1. In this problem, the learner has to decide which ad to display, amongst a given set that corresponds to the possible actions, on a web-page in order to maximize the probability that the Internet surfer clicks on it. This problem suscited many works in the community, see Lu et al. (2010),chun Wang et al. (2005), or Chakrabarti et al. (2008) for sparse examples.

Drug-allocation strategies. In this setting, a physician must find what is the best drug amongst many possible to administrate to a patient infected with some disease in order to cure him. This seemingly easy problem is actually old (Thompson (1933)) and is the one that led to the development Bandit game theory. Powerful algorithms have been designed to handle such problems thanks to the theory developed over years. In chapter 4, we consider natural extensions of this setting to adaptive viruses. Indeed, designing medical treatments for patients infected by adaptive viruses such that the Human Immunodeficiency Virus (HIV) is challenging due to their ability to mutate into new viral strains that become, with time, resistant to a specific drug Ernst et al. (2006).

Orientation at school. A less obvious application of reinforcement learning that is not currently addressed up to my knowledge is its use for orientation at school. In this problem, the goal is to find the subjects that maximize the probability of success of a student. This not only depends on marks and social environment, but also on the motivation of a student for a specific subject as well as on hidden variables. This goal is quite challenging for at least three reasons. First, the subjects typically evolve from one year to another one: a subject that is taught at a low level at school, e.g. maths at elementary school, is generally very low correlated with what is taught at a much higher level, e.g. maths at university. Thus, it is difficult to have a long term prediction. Also, the skills of the individual evolve with time, previous experience and training, personal events, etc. Finally, at the beginning of school, a student has access to various subjects, yet taught at a very low level. But after a student has discarded some subject by making choices, at high school or university, he/she has generally no easy opportunity to train again in another domain.

Thus we only have an approximate evaluation of the future success or failure of one student. But there is hope: One one hand, there are not so many possible choices - about a few hundreds for all possible outputs of the education system. On the other hand, for one student, we can have precise evolution of his success thanks to marks (but we also have to take into account other variables, like personal motivation). Moreover, there is a huge number of students that can give direct or indirect feedback after school about their degree of happiness or success. Thus we can probably exploit this three remarks to provide, thanks to machine learning, better orientation advices to students as well as precise modifications of the education system to improve success at school and right after school.

Economy and politics. Finally, let us consider the problem of ruling a state with the help of reinforcement learning. Thus, in this problem, we are the government. Our goal, from a country-wise perspective, is to manage the resources of all the country while maximizing the "happiness" of people.

By managing the resources, e we will consider that we have a given budget each year, and that we are allowed to modify the resources allocated to each class of business. We can use either a coarse notion of class, like "culture", "education','industry', etc. or we can use a very fine notion class, like "companies of more than 15 people that produce vehicles and respect the chart number 12345 on eco-friendly production line". The two extreme situations are the following one: on the one hand we consider just 1 class corresponding to all companies, institutes and organisms, in which case we do not manage resources at all. On the other hand we consider that each company is one class, which results in something like 10^5 or 10^6 many classes for a standard country. In this case, we control the resources of all companies. Of course between these two extremes, there is a trade off, and specific levels of details that enable to maximize the long term happiness of people and resources of the country. Finding an optimal categorization for resource allocation is not easy and can benefit from machine learning.

Now about the happiness, we consider that each individual has a specific objective - that may be not explicit - and that the more he/she succeeds, the happier. Note that the goal of a specific individual may be very different from the goal of the government, and of other individuals. The goal of two individuals can be for instance incoherent, which means that there exists no way to satisfy both. But although the goal of a specific individual is not known, we can have access to a rough estimation of the degree of happiness of people, simply by frequently asking to a small sample of the population how well one likes his/her job, salary, life, etc.

Due to the poor information we have, the possibility that people may change their goal, and the fact that the actions we take (allocation of resources) have only indirect effect on people, this task is highly difficult. Yet it seems that machine learning can provide important insights and be beneficial to this challenging problem. Up to my knowledge, I am not aware of the use of machine learning in that way for any current government.

Some challenges of reinforcement learning

In order to give more motivation to this thesis, we now consider the following possible artificial intelligence system - this is only one example amongst many. It already addresses a variety of challenging questions from the general agenda of Artificial Intelligence that goes far beyond what can answer up-to-date research and enables us to put reinforcement learning in context, enlightening some of, what I believe are, the major future challenges in the field.

Get observations from sensors. Let us consider an agent evolving in some real or virtual world. This agent has a bunch of *sensors*, like for instance sensors for temperature or pressure on your skin, one cell of your retina, etc. We will also consider without loss of generality that the sensing frequency is upper bounded, and thus that the time $t = 1, 2, ...$ is discrete. Each sensor $s \in \mathcal{S}$ outputs at time $t \in \mathbb{N}$ a real value $o_{s,t}$ that we assume to be in $[0, 1]$, and that we call an observation. Now, since at some time step t_0, we may not be able to memorize all past observations $(o_{s,t})_{s\in S, t\leqslant t_0}$, we focus on the most recent part, say the last τ observation steps, and compress the remaining part. We thus introduce the object $O_{t_0} = (o_{s,t})_{s\in S, t_0-\tau+1\leqslant t\leqslant t_0}$ that we call the matrix of observations at time t_0, and consists of all the observations from the time step $t = t_0 - \tau + 1$ to the current time step t_0, and the object $C_{t_0-\tau}$ that consists of the compression of the past observations not received after time step $t_0 - \tau$. At the next time step $t_0 + 1$, the observation matrix O_{t_0+1} consists of almost the same elements of O_{t_0}: it contains all the observations from time step $t_0 - \tau + 2$ to t_0, only the observations at time step $t_0 - \tau + 1$ are dropped and observations at time step $t_0 + 1$ are added. Now our goal will be to apply reinforcement learning, thus we need for that purpose states, actions, and rewards.

Build states from observations. We begin by identifying "states". A good property for a state is that it should not change if the agent does not do any action. Unfortunately, the observations may be subject to modifications independent from the learner actions, at any time step. One possible way to build states is to separate the observations into two parts, one part that is almost fixed when the learner does not perform any action, and a second one that evolves according to some unknown dynamic and that we will consider as a perturbation. For instance, consider you look at some scene, where someone is walking. Then the moving person will be considered as a perturbation, and you may not want to consider its complex behavior to define your current state, but you will prefer to define your state by the non-moving part, i.e. the background. This tells you for instance where you are. One way to perform such a decomposition is to apply a technique known as *low-rank matrix decomposition* (see Candés et al. (2009)) to the observation matrix O_{t_0}. This composition results in two objects, one with "low-rank" that roughly corresponds to the non moving part, we write it S_{t_0}, and a second perturbation part, that we write P_{t_0}. Note that provided we do not perform any action, one may also want to consider that the data corresponding to S_{t_0} for various t_0 is generated according to some underlying process in a similar way - we say identically and independently distributed or i.i.d. - and thus apply a universal code encoder that generates a symbol s_{t_0} that is a compressed representation of the observations and thus can play the role of an internal state. Note that in practice, it may be better to consider the data is generated by different i.i.d. signals mixed together, and thus try to learn a partition of the observations into several i.i.d signals that is the most effective in terms of coding theory. On the other, the perturbation part P_{t_0} can be seen hand as a complex process, that may be interesting to understand, but evolves according to its own complex dynamic. Thus it is natural to consider this perturbation as something "else", different than states. We can say

this defines generally an other agent.

Learn transitions between states. Now that we have a notion of states, we need to understand the way we jump from one state to another one when we perform some action. In other terms, we need to learn the transition probabilities of the state structure. Indeed, performing some action will create a modification of the signal received, and thus of S_t for some $t > t_0$. Of course this modification may also appear in the perturbation signal P_t for some $t > t_0$ as well, but it does not concerns our states then. We can learn the transition probabilities using standard reinforcement learning algorithms, provided the states are built such that they have some (Markov) property. Note that there may be in practice a possible delay between the time when the decision to perform some action is taken and the time when it is really performed, and also some so-called *trembling hand* effect which is an additional difficulty that is seldom considered from a theoretical perspective. A trembling hand effect is just the fact that we choose one action but instead another one is performed.

Identify reward signals. So far, we have not talked about rewards. But in order to apply the reinforcement learning theory and algorithms, we need to identify some reward signal. Fortunately, there are actually many ways to identify such a signal. We present here three ways that enable to define plenty of possible reward functions.

One direct way in order to identify a reward function is to look at the observations received from sensors and consider that one or several of our sensors provide(s) us with a reward signal. Our goal is thus to identify such a signal, since a reward signal has some specific structure. For instance, let us consider we have a set of energy sensors $E \subset \mathcal{S}$, and say that each sensor $e \in E$ measures the energy level of some source of energy like a battery. Now consider that when we perform one action a we use some energy corresponding to the source measured by e, and thus e decays; when the energy level is 0, we can no longer perform action a. Each time we play this action, the decay is roughly the same, and thus it should be able to automatically identify this signal as being a negative reward. Identifying which sensor(s) provides us with rewards seem not to be a well studied problem. It seems there are ways to infer detect signals satisfying "good" reward properties, although this is a bit tedious since there is not always a ground truth. Now once a reward signal is identified, one may want to use this knowledge and thus redefine states accordingly, which shows that reward identification problem creates a variety of not theoretically-dressed interesting questions.

Another way to identify rewards is to look at the perturbation signal, for we may consider this corresponds to the observation of some other agent that evolves in the world, and tries to achieve some goal. The problem that consists in inferring a reward function from the observation of an agent has received some interest, and is known as Inverse reinforcement learning (see Ng and Russell (2000), Ramachandran and Amir (2007), Abdeslam et al. (2011)). However this a really challenging problem that is not completely formalized and thus known results are pretty weak.

Finally, there is at least a third way to define a reward function. The idea is that

some actions may enable to better understand the world and thus build more accurate and compressed internal representation than others. This compression progress can be quantified and measured, and thus can be used as a reward signal. It corresponds to the notion of intrinsically motivated rewards that is for instance developed in Schmidhuber (2009).

Solve a reinforcement learning problem. Now that we have identified states, transitions, and that we can consider some reward function, we can learn how to act almost optimally with respect to this reward function and this representation of the world.

More precisely, we can use the formalism of Markov Decision Process (MDP) and then use all the known literature of reinforcement learning to solve this problem (see for instance Auer and Ortner (2006), Boyan (1999), Lazaric et al. (2010a), Scherrer (2010)). Note, however, that we here have to deal with an approximate MDP due to the identification of states and rewards; fortunately, extending the notion MDP to other setting is a problem that receives increasing interest (see for instance Chakraborty and Stone (2010)).

Target many goals. Finally, in this red-line example, as it is the case in practice, we have the choice to target many different goals, i.e. problems defined by one reward function, either inferred from an other agent, identified from our sensors, or defined internally. Thus we need to have a decision procedure to select which goal to follow, at a high level. For instance, some problems may be learned quickly, others may more difficult but may also help to solve a lot of other problems, and we generally do not know in advance the intrinsic complexity of a problem. Thus this question is challenging (this is actually A.I.) and goes beyond the scope of classical reinforcement learning.

Of course, one way to address this question would be to reduce the problem to reinforcement learning by introducing an additional high level reward, but it is difficult to say one kind of meta-reward is well-suited and thus, this is now a philosophical question. From a philosophical point of view, we may consider at least three different (very) high level goals.

- "I do not want to die". Due to energy consumption, the agent has to act in order to get energy and thus avoid that its energy falls to zero. Note that since performing some action that is immediately energy consuming may increase its lifetime on the long run, this goal may lead to not trivial behaviors. The first naive goal is thus to maximize the lifetime of the agent.

- "I want the world to remember me". Since actions have effect in the future, and some may have effect in the long run, this means that the lifetime of the learner is lower than the time during which it has an impact on the environment, that we call the "impact" time. Thus, it makes sense to try to maximize its impact time. This is the second goal. Note that trying to have a long life may be an interesting sub-goal in order to have enough time to perform high-impact actions. Note also that, if we consider the set of all actors, i.e. all agents that may act on the environment, an other natural sub-goal in order to maximize its impact time is to perform some action that directly makes the

other agents modify their behavior. Indeed, this way, the initial agent still has some indirect impact on the environment.

- "I want to promote life". Since there is a priori no reason why the agent should be considered differently than other agents, it makes sense to consider the set of all agent as being one big agent including the agent itself. In this case, the natural goal of the agent is to maximize the life expectancy of the big agent, i.e. act in order to make sure that there will be still agents in the future able to act on the environment. For instance, this can be done by performing actions that may bring energy to other agents in the future. This is somehow linked to the first two goals but we here just care about the big agent, not really about which part of the agent survives longer.

Which one of these three goal is "better" is highly debatable, but it is clear that none of those are easily achievable, even for human beings, and that we will surely have to wait one or several other decades in order to be able to formalize these questions for machines. Thus we are not going to address this question in this thesis, nor the many others that surround this general example, but we instead focus on the reinforcement learning part for which a lot has to be understood, and leave them for a future work.

Contents

Roadmap

Before starting the main matter of this Ph.D. dissertation, we now propose a general roadmap in order to help the impatient reader. It is summarized in figure 4 below.

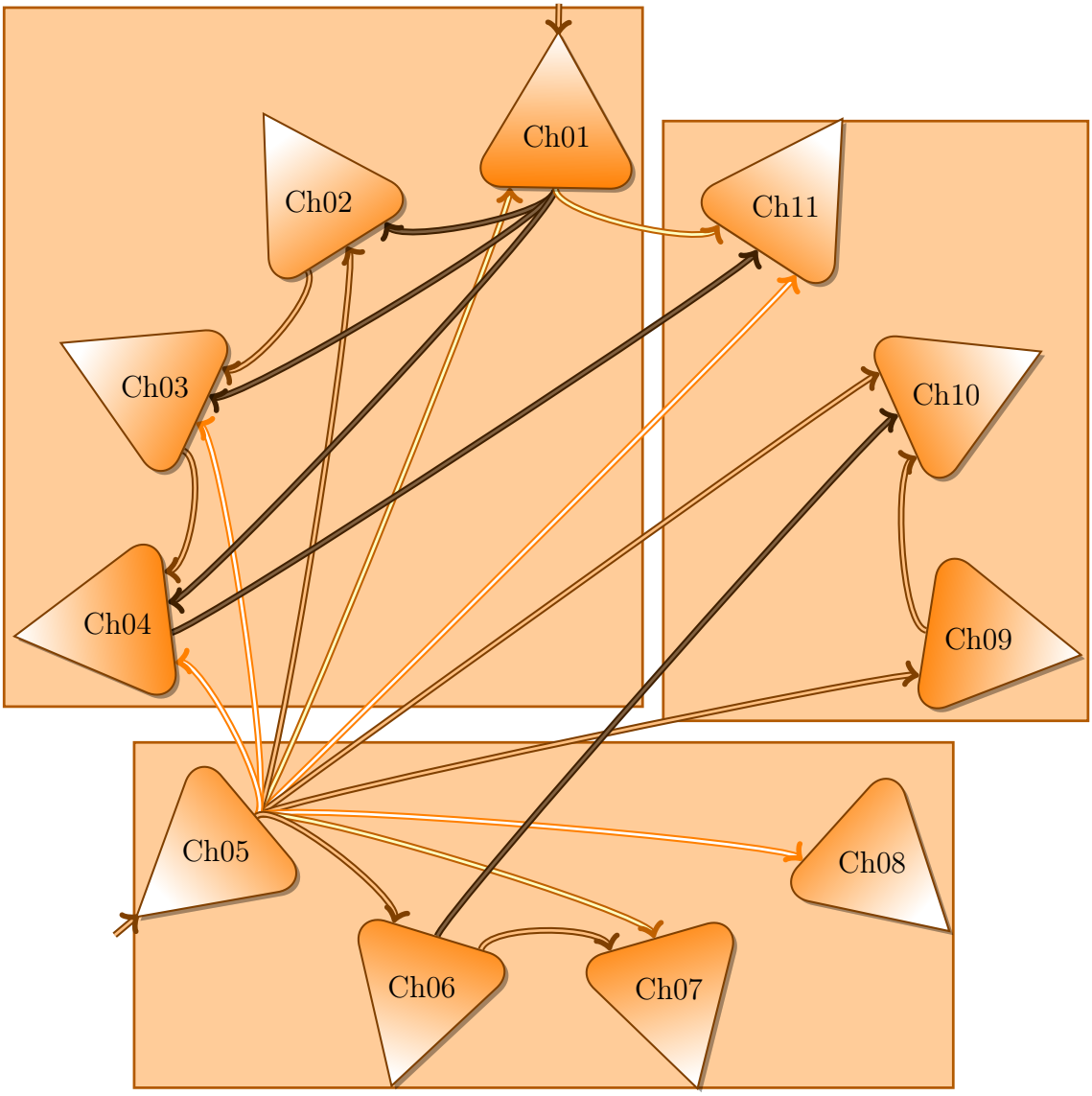


Figure 4: General reading roadmap for this Ph.D dissertation

Figure 4 reads as follows. The entry points of these dissertation are indicated by the two input arrows. These are (obviously) chapter 1 and also chapter 5 as they correspond to two

general surveys that present the background material needed to understand the setup and context of the questions addressed in this Ph.D. dissertation.

Now the chapters of this dissertation are gathered into three main parts. Each part is indicated by a rectangular box on the figure. In each part, we suggest starting to read the first chapter of the part - these are respectively chapter 1, 5 and 9 - and then follow the proposed natural ordering of the chapters. Since one may also want to jump directly to a specific contribution, and also for clarity purpose, we indicate the structure of dependency between chapters via wires.

The meaning of a wire starting from chapter n and going to chapter m is that chapter m uses some concepts that are presented or explained with more details in chapter n. Moreover, *the darker the wire, the stronger the dependency*. Thus for instance according to figure 4, it is strongly advised to read chapter 1 before reading chapter 4, and similarly chapter 6 before chapter 10. On the other hand, chapter 3 makes use of some concepts of chapter 5, but only very scarcely.

Interestingly, one can see from the figure that the concepts of chapter 5 are used in all other chapters, making this chapter seemingly the most important. Actually this can be understood since this short chapter gathers most of the statistical tools used here and there in the other chapters. It is thus important because of technical details, but actually does not provide as much intuition about the rest of this dissertation as chapter 1 for instance. This is also the reason why the wires going out from chapter 5 are lighter.

Finally, we do not report in this figure the important fact that part III can be seen as the result of combining the concepts that underlie part I and part II.

Contributions.

Cette section est l'occasion de présenter un résumé des différentes contributions de cette thèse dans les domaines de recherche des jeux de bandit, de l'apprentissage statistique et de l'apprentissage par renforcement.

Partie I.

La première partie se concentre sur le cadre des Bandits, cadre à la fois fondamental pour comprendre l'apprentissage par renforcement et également intéressant en lui-même.

Le monde des bandits: Exploration et Exploitation.

Le chapitre 1 présente une revue (non exhaustive) de la littérature incroyablement vaste concernant les bandits. Nous présentons ensuite trois contributions à la théorie des Bandits, depuis le cadre le plus standard des bandits dits stochastiques au cadre plus ambitieux des bandits adverses, jusqu'au cadre difficile des bandits adaptatifs que nous introduisons dans le chapitre 4.

Chapitre 2: Bandits à plusieurs bras utilisant la divergence de Kullback-Leibler. Le chapitre 2 concerne le problème des bandits à plusieurs bras dits stochastiques. Son but est de fournir une analyse, à la fois en temps fini et dépendante de la distribution de lois sur chaque bras, d'un algorithme que nous prouvons être optimal en un certain sens, ce qui répond à une question ouverte il y a longtemps par les travaux de Lai and Robbins (1985) et plus tard de Burnetas and Katehakis (1996).

Nous analysons le cadre des bandits à plusieurs bras stochastiques et plus précisément l'écart qui apparaît entre, d'une part, les bornes asymptotiques qui furent prouvées dans Lai and Robbins (1985) puis dans Burnetas and Katehakis (1996) et qui font apparaître la divergence de Kullback-Leibler entre les distributions des bras, et, d'autre part, les bornes non-asymptotiques qui furent prouvées pour des algorithmes du type UCB dans Auer et al. (2002) puis plus tard dans Audibert et al. (2009), Audibert and Bubeck (2010), Auer and Ortner (2010) etc. et qui font apparaître seulement le premier ou le deuxième moment des distributions de chaque bras.

Les premières bornes atteignent la borne inférieure asymptotique de performance, mais sont seulement asymptotiques et ne concernent que des classes de distributions spécifiques (certaines distributions paramétrées de dimension finie), tandis que les secondes sont non-asymptotiques et concernent des distributions arbitraires (de support inclus dans $[0, 1]$), mais

malheureusement ne permettent pas d'atteindre la borne inférieure asymptotique. De plus, comme cela est mentionné dans les études expérimentales menées par Filippi (2010) ou Honda and Takemura (2010a), les algorithmes utilisant la divergence de Kullback-Leibler atteignent des performances significativement meilleures que celles des algorithmes du type UCB.

Nous comblons partiellement cet écart en étudiant un algorithme utilisant la divergence de Kullback-Leibler pour le problème des bandits stochastiques à plusieurs bras dans le cas de distributions à support fini (i.e. avec un nombre fini d'atomes), dont le regret asymptotique correspond à la borne inférieure de Burnetas and Katehakis (1996), et dont on fournit une analyse non asymptotique.

Chapitre 3: Algorithmes de bandit pour l'apprentissage en ligne dans des environnements adverses Lipschitz.

Le chapitre 3 s'attaque au problème de bandit dit adverse, en information complète, lorsque un grand ensemble de bras est considéré, ainsi qu'à ses applications à l'apprentissage en ligne. Nous dérivons, sous certaines hypothèses géométriques et topologiques faibles, des bornes de performance pour un algorithme ainsi qu'un schéma d'approximation numérique efficace utilisant des techniques de Population Monte-Carlo.

Puisque le cadre de l'information complète permet de traiter un grand nombre de bras, nous abandonnons l'hypothèse d'un ensemble fini de bras $\mathcal{A}$ au profit d'un sous-ensemble de $\mathbb{R}^d$, ce qui permet d'appliquer les bandits au problème de l'apprentissage en ligne lorsque l'environnement est un adversaire. Pour de grands ensembles, il faut supposer une certaine régularité des fonctions de récompense afin de contrôler le terme regret. La difficulté principale, cependant, est de fournir une implémentation numérique efficace, ce qui requiert généralement de faire des approximations de l'algorithme théorique naturel.

Ici nous considérons le problème de l'apprentissage en ligne face à un environnement adverse lorsque les fonctions de récompense choisies par l'adversaire sont supposées être Lipschitz. Ce cadre étend des travaux précédents sur l'apprentissage en ligne dans un cadre linéaire (cf. Dani et al. (2008a), Abernethy et al. (2008b), Cesa-Bianchi and Lugosi (2009), Kakade et al. (2008)) ou convexe (cf. Zinkevich (2003), Hazan et al. (2006)). Nous étudions une classe d'algorithmes dont le regret cumulé est borné supérieurement par $\tilde{O}(\sqrt{dT\ln(\lambda)})$, où d est la dimension de l'espace de recherche, T est l'horizon temporel, et λ est la constante de Lipschitz.

Nous discutons la question importante de fournir une implémentation numérique efficace et utilisons des méthodes particulaires dans ce but. Notons qu'un travail récent de Narayanan and Rakhlin (2010) montre qu'avec l'hypothèse plus forte de fonctions de récompenses convexes, la méthode d'échantillonnage particulaire peut être légèrement simplifiée (en utilisant un échantillonneur de Gibbs pour les mesures log-concaves) en conservant un algorithme à la fois sain et numérique efficace. Les applications que l'on considère regroupent des problèmes d'apprentissage supervisé en ligne en information complète ainsi qu'en information partielle (bandits), pour une large classe de régresseur/classificateurs comme par exemples les réseaux de neurones.

Chapitre 4: Bandits adaptatifs: vers la meilleure strategie dépendant de l'historique
Dans le chapitre 4, nous considérons le cas d'un problème de bandits multi-armé en information partielle lorsque l'environnement est possiblement adaptatif à l'apprenant, sans être forcément le pire environnement possible. Nous introduisons une définition de regret permettant de capturer une telle notion, et nous montrons comment un algorithme peut bénéficier de cette notion et ainsi être adaptatif en un certain sens à la complexité de l'adversaire.

La raison pour ne pas considérer uniquement le pire adversaire est qu'en pratique, un algorithme n'affrontera pas nécessairement un tel adversaire, et qu'un algorithme conçu uniquement pour le pire cas ne profitera pas nécessairement des faiblesses de l'adversaire. Concevoir des algorithmes adaptatifs à la faiblesse de l'adversaire est un défi.

Nous introduisons ici un modèle de contraintes Θ, fondé sur des classes d'équivalence de l'historique commun (i.e. l'information partagée par le joueur et l'adversaire), qui définit deux scenarii d'apprentissage: (1) L'adversaire est contraint, i.e. il fournit des récompenses qui sont des fonctions stochastiques des classes d'équivalence définies par un modèle $\theta^* \in \Theta$, et l'on mesure le regret par rapport à la meilleure stratégie dépendante de l'historique. (2) L'adversaire est arbitraire, et l'on mesure le regret par rapport à la meilleure stratégie parmi toutes les fonctions allant des classes vers les actions (i.e. la meilleure stratégie basée sur les classes d'historique) pour le meilleur modèle dans Θ. Ceci permet de considérer des modèles d'adversaires (cas 1) ou de stratégies (cas 2) incluant ceux à mémoire finie, périodiques, les bandits stochastiques standards et bien d'autres situations.

Lorsque $\Theta = \{\theta\}$, i.e. un seul modèle est considéré, nous dérivons des algorithmes numériquement efficaces dont le regret (au temps T) est finement borné par $\tilde{O}(\sqrt{TAC})$, où C est le nombre de classes de θ. A présent, lorsque plusieurs modèles sont disponibles, tous les algorithmes connus atteignant une bonne borne $O(\sqrt{T})$ sont malheureusement non numériquement efficaces et s'étendent difficilement à un grand nombre de modèles $|\Theta|$. Notre contribution ici est de fournir des algorithmes numériquement efficace ayant un regret borné par $T^{2/3}C^{1/3}\log(|\Theta|)^{1/2}$.

Partie II.

Après la première partie consacrée à quelques variations autour du problème des bandits, que l'on peut voir comme un problème purement en ligne en comparaison du problème général de l'apprentissage par renforcement, nous étudions dans une deuxième partie quelques questions importantes liées à l'apprentissage par lot, c'est à dire lorsque on nous donne un ensemble et non un flux de données.

L'apprentissage par lot: Randomisation et Échantillonnage.

Le chapitre 5 présente un aperçu général des outils de théorie statistique que l'on regroupe ici pour des raisons de clarté, puisque la plupart des théorèmes qui y sont présentés sont utilisés ici et là dans ce manuscrit de thèse.

Chapitre 6: Régression linéaire utilisant les projections aléatoires.
Dans le chapitre 6, on s'intéresse à l'utilisation de matrices aléatoires dans le cadre de la régression en design aléatoire. Si les outils nécessaires pour établir les bornes de performance des estimateurs proposés ont été popularisés assez récemment en raison des nombreux développements pratiques et théoriques auxquels a mené le sujet des matrices aléatoires au cours des dernières années, il est intéressant de réaliser que l'idée d'utiliser les projections aléatoires, ou les représentations aléatoires telles qu'elles sont nommées dans Sutton (1996), est déjà ancienne dans des domaines davantage appliqués, tels que la robotique ou la synthèse de texture par exemple, ce qui donne plus de motivation pour les comprendre. Par exemple, Richard Sutton étudiait expérimentalement les effets de la randomisation dans les réseaux de neurones déjà dans Sutton and Whitehead (1993), et mentionne que le Perceptron de Rosenblatt en 1962 était originellement utilisé avec une couche initiale de randomisation afin d'améliorer les performances.

Nous étudions une méthode de régression qui construit, à partir d'un espace de fonction $\mathcal{F}$ donné de grande dimension (possiblement infinie), par exemple $L_2([0,1]^d;\mathbb{R})$, un sous espace $\mathcal{G}_P \subset \mathcal{F}$ de dimension finie P généré aléatoirement. $\mathcal{G}_P$ est défini comme l'espace linéaire engendré par P éléments aléatoires, eux même obtenus par combinaison linéaire de fonctions de base de $\mathcal{F}$ pondérées par des coefficients aléatoires gaussiens iid. Nous présentons une motivation pratique pour utiliser cette approche, détaillons le lien que partagent ces projections aléatoires avec la théorie des RKHS et des objets Gaussiens, établissons, en design déterministe et également aléatoire, des bornes sur l'erreur d'approximation lorsque l'on cherche la meilleure fonction de régression dans $\mathcal{G}_P$ au lieu de $\mathcal{F}$, et dérivons des bornes d'excès de risque pour un algorithme de régression spécifique (régression par moindre carrés dans l'espace $\mathcal{G}_P$). Ce papier mets l'accent sur la motivation pour étudier de telles méthodes, ainsi l'analyse développée reste simple à des fins de meilleure explicitation, et laisse la place à de futures extensions et améliorations.

Chapitre 7: Échantillonnage Brownien pour la reconstruction de fonctions parcimonieuses.
Dans le chapitre 7, nous considérons une autre utilisation des matrices aléatoires de manière plus traditionnelle, en lien avec le problème de reconstruction d'une fonction parcimonieuse. Spécifiquement, nous montrons comment l'utilisation d'opérateurs d'intégration aléatoires permet de relâcher des hypothèses classiques de (quasi) orthogonalité du dictionnaire sous-jacent, en transformant le problème de reconstruction en un simple problème d'intégration.

Le chapitre précédent montrait le bénéfice qu'il y a à utiliser les matrices aléatoires pour s'attaquer au problème de prédire aussi bien qu'une fonction cible inconnue. Ici, nous nous intéressons au problème de reconstruction, où le but est de reconstruire le paramètre de décomposition de la fonction inconnue. Ce problème est généralement plus difficile puisque reconstruire ce paramètre entraîne naturellement une faible erreur de prédiction par rapport à la fonction cible.

Plus précisément, nous considérons le problème de reconstruction du paramètre de dé-

composition $\alpha \in \mathbb{R}^K$ d'une fonction f supposée parcimonieuse dans une famille de fonctions connue $\{\varphi_k\}_{1 \leqslant k \leqslant K}$ (i.e. le nombre de composantes non nulles du vecteur α est petit par rapport au nombre total de composantes K) , à partir d'évaluations bruitées de f sur un ensemble bien choisi de points d'échantillonnage. Nous introduisons un processus de randomisation supplémentaire, appelé Brownian sensing, reposant sur le calcul d'intégrales stochastiques, ce qui génère une matrice d'échantillonnage Gaussienne pour laquelle on démontre de bonnes propriétés de reconstruction, indépendamment du nombre de point N et lorsque les fonctions de bases sont arbitrairement non orthogonales. Sous l'hypothèse que f est Hölder d'exposant au moins $1/2$, on propose un estimateur $\widehat{\alpha}$ du paramètre tel que $\|\alpha - \widehat{\alpha}\|_2 = O(\|\eta\|_2/\sqrt{N})$, où η est le bruit d'observation. La méthode utilise un ensemble de points uniformément distribués selon une courbe de dimension un sélectionnée en fonction des fonctions de base. Nous rapportons des résultats d'expérience numérique qui illustrent notre méthode.

Chapitre 8: Apprentissage multi-vue: Complexité versus Consensus.
Enfin dans le chapitre 8, bien qu'un peu déconnecté du reste de ce manuscrit, nous analysons la complexité de Rademacher d'un problème dit d'apprentissage multi-vue.

Nous considérons le problème de la classification multi-vue semi-supervisée, où chaque vue est supposée correspondre à un Espace de Hilbert à Noyau Reproduisant. Nous étudions un algorithme utilisant des méthodes de co-régularisation utilisant des termes de pénalité supplémentaires reflétant des propriétés de continuité et de consensus entre les vues. Ce travail fournit à la fois une borne supérieure et inférieure explicite sur la complexité de Rademacher de la classe d'apprenant correspondant, pour un nombre arbitraire de vues. Nous montrons également le comportement asymptotique des bornes lorsque le terme de co-régularisation augmente, rendant ainsi explicite la dépendance entre la *consistance* entre vues et la *réduction* de l'espace de recherche. Nous appliquons cet algorithme à plusieurs exemples jouets incluant un nouvel exemple non trivial. Enfin, nous prenons parti pour une méthode de sélection de paramètres basée sur une notion de stabilité inspirée par le clustering et des arguments de localisation. Nous fournissons des bornes explicites sur la variance de la classe et proposons un algorithme de sélection.

Partie III.

Dans la dernière partie, nous combinons le monde des bandits et de l'apprentissage par lot, afin de s'attaquer à des problèmes d'apprentissage par renforcement.

Vers le monde réel(?): Modélisation et Planification.

Les trois derniers chapitres traitent trois questions différentes en apprentissage par renforcement, liés aux Processus Décisionnels de Markov (PDM).

Chapitre 9: Analyse en temps fini de l'algorithme de minimisation du résidu de Bellman.
Dans le chapitre 9, on analyse la minimisation du résidu de Bellman: il s'agit d'un algorithme naturel dans le cadre des PDM actualisés lorsque l'on a recours à un modèle génératif, i.e. que l'on peut échantillonner n'importe quand une action à partir de n'importe quel état, par opposition au cadre où on ne peut qu'échantillonner une action à partir de l'état courant.

Nous considérons l'approche par minimisation du résidu de Bellman pour résoudre des problèmes décisionnels de Markov actualisés, où l'on suppose qu'un modèle génératif de la dynamique et des récompense est disponible. A chaque étape d'itération sur la politique, une approximation de la fonction valeur de la politique courante est obtenue en minimisant un résidu de Bellman empirique défini sur un ensemble de n états tirés de manière i.i.d à partir d'une distribution μ, des récompenses immédiates et des états suivants échantillonnés à partir du modèle. Notre résultat principal est une borne de généralisation pour le résidu de Bellman dans des espaces d'approximation linéaires. En particulier, nous démontrons que le résidu de Bellman empirique approche le vrai résidu de Bellman (quadratique) en norme-μ avec une vitesse en $O(1/\sqrt{n})$. Ce résultat implique que minimiser le résidu de Bellman empirique est en effet une approche bien fondée pour la minimisation du vrai résidu de Bellman, ce qui garantit une bonne approximation de la fonction valeur pour chaque politique. Enfin, nous dérivons des bornes de performance pour l'algorithme d'itération sur les politiques approché résultant de cette méthode, en terme du nombre d'échantillons n et d'une mesure de complexité indiquant la capacité de l'espace de fonctions à approcher les fonctions valeurs successives.

Chapitre 10: Différences temporelles par moindre carrés avec projections aléatoires.
Dans le chapitre 10, nous analysons une version d'un algorithme appelé différences temporelles par moindre carrés, où l'on utilise des projections aléatoires telles que présentées dans le chapitre 6, dans le but de tirer parti de la réduction de dimension. Cet algorithme est conçu pour des PDM actualisés lorsque l'on n'a pas accès à un modèle génératif et donc que l'on est donc forcé d'échantillonner depuis l'état courant, en suivant une seule trajectoire. Il est intéressant de constater que, du point de vue de la théorie statistique de l'apprentissage, le problème d'estimation de la fonction valeur correspondant à cet algorithme peut être vu comme un problème de régression en design Markovien, où la fonction cible ne peut être échantillonnée directement comme d'ordinaire, mais est définie au contraire comme point fixe de l'opérateur de Bellman que l'on doit estimer.

Nous considérons le problème d'apprentissage par renforcement dans des espaces de grande dimension lorsque le nombre de fonctions de base est plus grand que le nombre d'échantillons. En particulier, nous étudions l'algorithme de différence temporelle par moindre carrés (LSTD) lorsque un espace de petite dimension est généré par projection aléatoire à partir d'un espace de grande dimension. Nous fournissons une analyse théorique complète de l'algorithme LSTD avec projections aléatoires et dérivons des bornes de performance pour cet

algorithme. Nous montrons également comment l'erreur de LSTD avec projections aléatoires se propage à travers les itérations d'un algorithme d'itération sur la politique et fournissons une borne de performance pour l'algorithme d'itération sur les politiques par moindre carré (LSPI) correspondant.

Chapitre 11: Sélectionner la représentation des états en apprentissage par renforcement.
Enfin le chapitre 11 pose quelques pierres vers la solution au problème important de sélectionner un modèle d'états pour l'apprentissage par renforcement. En effet, en pratique, il peut être difficile de définir une bonne notion d'états, et il peut donc y avoir ainsi plusieurs modélisations possibles. Nous construisons notre analyse au-dessus de l'algorithme UCRL2 conçu pour des PDMs non actualisés, et considérons un cadre philosophiquement relié au chapitre 4 qui traite la question difficile des bandits adaptatifs.

Dans ce chapitre, plusieurs modèles (fonctions allant des observations passées vers un ensemble fini) d'observations sont donnés, et l'on sait que pour au moins un de ces modèles, la dynamique résultante est en effet Markovienne. Sans connaître ni lequel de ces modèles est le bon, ni quelles sont les caractéristiques probabilistes du PDM résultant, le but est d'obtenir autant de récompenses que la politique optimale du bon modèle (ou du meilleur modèle, s'il y en a plusieurs). Nous proposons un algorithme qui atteint cet objectif, avec un regret de l'ordre $T^{2/3}$ où T est l'horizon temporel.

Part I

The World of Bandits: Exploration and Exploitation.

In this first part, we focus on the setting of Bandits that is both fundamental in order to understand Reinforcement learning and interesting by itself.

Why the setting of Bandits is important. Artificial Intelligence is interested in designing artificial agents that evolve in an environment. These agents sequentially observe the environment through sensors, learn, adapt, and take decisions, i.e. output actions.

The setting of *Reinforcement Learning* is a way to formalize what we mean by agent, by making decisions, and what the agent should learn; in this setting, it is assumed that the agent receives a reward to the action he/she has output, that is a real value. The reward measures the quality of this action. The basic goal of the agent is to learn how to output actions so as to maximize the sum of received rewards. Note that decisions are taken sequentially at each time step, and the environment and reward functions may evolve with time.

An informal definition of the multi-armed Bandit setting is that a learning agent is facing a bandit, i.e. a casino slot machine with a finite set of arms $\mathcal{A}$. At each time step t the learner can pull one arm $a_t \in \mathcal{A}$, and with each arm $a \in \mathcal{A}$ is associated an unknown and fixed probability law ν_a on the output rewards. Thus the learner receives one reward distributed according to ν_{a_t} after he/she chooses a_t. The game is repeated T times and the goal is to maximize the sum of received rewards up to time T.

The first reason to study bandits is that it can be seen as a base stone to understand an important notion in reinforcement learning known as Markov Decision Processes (MDP). In a nutshell (see Part III for more precise definition), a MDP models the environment thanks to states, and from any state, when an action is chosen, we move to another state according to a generally not known transition kernel. The reward function is a function of the states and actions, which means that the agent not only has to learn which action to output, but better learn a strategy, i.e. a decision rule that defines which action to output in which state. The goal is now to output a strategy that is optimal, i.e. that enables to receive the maximal sum of rewards whatever the initial state. Now a bandit can be seen as a MDP with only one looping state, thus studying this first problem in details, which is already not that easy, is an important tool towards solving the MDP problem: one idea is to decompose a MDP into different bandit problems and then combine them in a careful way.

The second reason to study bandits is that even more than fifty years after it was formally introduced by Robbins (1952), there are still many fruitful extensions one may consider to the original setting for practical purpose, together with many practical and theoretical opened questions that go much beyond the setting of reinforcement learning. Actually one can say that a real theory of bandits has emerged from these developments, with its own difficult questions and its own real-life applications.

Contributions. We present in Chapter 1 a general (not exhaustive) survey of the incredibly huge literature about bandits. Then we present three contributions to the bandit theory, from the most standard setting of so-called stochastic bandits to the more challenging setting of

adversarial bandits and then to the difficult case of adaptive bandits that we introduce in the last chapter.

Chapter 2 is about the stochastic multi-armed bandit problem and aims at providing a finite-time distribution-dependent analysis of an algorithm that we prove to be optimal in some sense, which fills a gap opened long ago in Lai and Robbins (1985) and later in Burnetas and Katehakis (1996).

In Chapter 3, we are interested in the so-called adversarial bandit problem when a large set of arms is considered, and in its application to online learning. We prove performance bounds for an algorithm under some weak geometrical and topological condition on the problem, together with a numerically efficient approximation scheme that uses Population Monte-Carlo technique.

Finally in Chapter 4, we consider the case of a bandit problem when the environment is considered to be possibly adaptive to the learner, but may be different from the worst possible environment. We introduce a definition of regret that enables to capture such a notion, and show how an algorithm can benefit from this notion and be adaptive in some sense to the complexity of the opponent.

Chapter 1

Multi-armed Bandit Games.

This introductory chapter is about a sequential decision problem called Bandits, named after the casino slots machines. We first introduce the setting and present a general overview of the standard results and algorithms, then we present many fruitful extensions of the initial setting together with pointers to the corresponding works. In the third part we consider the general class of decision makers known as Exponentially-weighted forecasters, for which we give some geometrical interpretation. Finally, we briefly discuss the limitations of bandits and show the need to consider other settings.

Contents

1 The standard stochastic multi-armed Bandit setting

In this section, we detail the original bandit problem, which enables to fix some notation, define the notion of regret and present the main algorithms that solve this problem.

Origin. The term bandit refers to the name of casino slot machines: the player uses a coin, then pulls the arm of the machine and receives a random amount of money (reward). Since there is only one arm (action), this is also called a one-armed bandit problem. The multi-armed bandit problem corresponds to the case when the player faces a machine with a finite number of arms, or equivalently a finite number of machines with one arm, and selects sequentially, by using one coin at each time step the arm with which he/she wants to play. Then the player receives the corresponding random reward and the goal is to earn as much money as possible.

Motivation. The historical motivation for this setting directly comes from medical trials, as introduced in Thompson (1933) for the comparison between two treatments, and then in Thompson (1935) for the more general case of finitely-many treatments. In this problem, there is a set $\mathcal{A}$ of drugs available in order to cure one specific disease. Patients come sequentially and it is assumed that each drug acts in the same way on each patient (i.i.d. assumption). The success of a drug on a patient is modeled by a Bernoulli random variable whose parameter depends only on the drug. Since each trial involves a human, we want to make as few mistakes as possible while focusing as soon as possible on the best drug.

1.1 Setting

More precisely, this problem has been formalized quite a long time ago by Robbins in Robbins (1952). In its original formulation, one considers a finite set $\mathcal{A}$ of A many arms. Each arm corresponds to a source of random variables independent and identically distributed (i.i.d.) according to an unknown probability distribution over the unit interval $[0,1]$.

The game is *sequential* and goes as follows: at each round $t \geqslant 1$, the player first picks an arm $A_t \in \mathcal{A}$ and then receives a stochastic payoff Y_t drawn at random according to ν_{A_t}, and only gets to see the payoff Y_t. We then define the cumulative reward of the forecaster up to time T to be $\sum_{t=1}^{T} Y_t$.

The goal of the forecaster is to maximize its expected cumulative reward up to the time horizon T. The forecaster may or may not know in advance the horizon. When the forecaster does not know in advance this horizon, we say the strategy is anytime.

Regret definition. For each arm $a \in \mathcal{A}$, we denote by μ_a the expectation of its associated distribution ν_a and we let $a^\star$ be any optimal arm, i.e., $a^\star \in \arg\max_{a \in \mathcal{A}} \mu_a$.
We write $\mu^\star$ as a short-hand notation for the largest expectation $\mu_{a^\star}$ and denote the gap of the expected payoff μ_a of an arm $a \in \mathcal{A}$ to $\mu^\star$ as $\Delta_a = \mu^\star - \mu_a$. In addition, the number of times each arm $a \in \mathcal{A}$ is pulled between the rounds 1 and T is referred to as $N_T(a)$,

$$N_T(a) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}} .$$

Stochastic multi-armed bandit
Unknown parameters: $|\mathcal{A}|$ unknown probability distributions on $[0,1]$
Known parameters: the set of arms $\mathcal{A}$, the number of rounds T
For each round $t = 1, 2, \dots, T$

(1) the learner chooses $A_t \in \mathcal{A}$ according to its own strategy.

(2) the learner incurs and observes the reward $Y_t \overset{i.i.d.}{\sim} \nu_{A_t}$ independently from the past given A_t.

Figure 1.1: The stochastic multi-armed bandit game.

Now the quality of a decision strategy is evaluated via the notion of regret that we define precisely now.

Definition 1.1 (*Expected regret*) *The expected regret, or just the regret, at round $T \geqslant 1$ is defined as*

$$R_T \overset{\text{def}}{=} \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} Y_t\right] = \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} \mu_{A_t}\right] = \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}\big[N_T(a)\big]\,, \tag{1.1}$$

where we used the tower rule for the first equality. Note that the expectation is with respect to the random draws of the rewards according to the distribution ν_{A_t} and also to the possible auxiliary randomization to which the decision-making strategy is resorting.

The meaning of this definition is that we measure the *regret* of the forecaster for not playing optimally. Indeed, the strategy of the oracle, i.e. when all the distributions are known, would be to constantly select the arm with the highest mean reward. Thus, in order to measure the performance of a specific forecaster, we compare its mean cumulative reward to the mean cumulative reward of the oracle strategy.

Lower bounds. It is important to know what kind of regret is possible to reach. For that purpose, we are interested in lower bounds on the expected regret for the best decision-maker. We consider two kind of bounds:

(1) Distribution-free lower bounds, i.e. bounds that do not make appear quantities specific to the law of the arms. They are important in order to known what are the best performances one can hope for in the worst case, or uniformly over the classes of distributions.

(2) Distribution-dependent lower bounds, i.e. bounds that makes appear quantities that depend on the law of of the arms. They are important in order to know what are the best performances one can hope for one bandit problem.

In the case of distribution-free lower bounds (also called minimax lower bound), we have the following result from Auer (2003).

Theorem 1.1 (*Minimax regret lower bound*) *Let sup represents the supremum taken over all stochastic bandits with support in* $[0,1]$ *and inf the infimum taken over all forecasters, then the following holds true:*

$$\inf \sup R_T \geqslant \frac{1}{20}\sqrt{TA}$$

Now in order to introduce distribution-dependent lower bounds, let us first remind some useful notion from information theory.

Kullback-Leibler divergence. We denote by $\mathcal{P}([0,1])$ the set of probability distributions over $[0,1]$. For two elements ν, $\kappa \in \mathcal{P}([0,1])$, we write $\nu \ll \kappa$ when ν is absolutely continuous with respect to κ and denote in this case by $\mathrm{d}\nu/\mathrm{d}\kappa$ the density of ν with respect to κ. We recall that the Kullback-Leibler divergence between ν and κ is defined as

$$\mathcal{K}(\nu,\kappa) = \begin{cases} \displaystyle\int_{[0,1]} \frac{\mathrm{d}\nu}{\mathrm{d}\kappa} \log \frac{\mathrm{d}\nu}{\mathrm{d}\kappa}\, \mathrm{d}\kappa & \text{if } \nu \ll \kappa; \\ +\infty & \text{otherwise.} \end{cases} \tag{1.2}$$

We first state the following Theorem, adapted from Lai and Robbins (1985) for the case when all the distributions are Bernoulli distributions:

Theorem 1.2 (*Regret lower bound for Bernoulli distributions*) *Let us consider a consistent forecaster, i.e. such that for any stochastic bandit, for any suboptimal arm a and any $\beta > 0$, $\mathbb{E}(N_T(a)) = o(T^\beta)$. Then for any stochastic bandit with Bernoulli distributions, all different from a Dirac distribution at 1, the following holds true:*

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geqslant \sum_{a:\Delta_a>0} \frac{\Delta_a}{\mathcal{K}(\mathcal{B}(\mu_a),\mathcal{B}(\mu^\star))}\,,$$

where $\mathcal{B}(p)$ stands for a Bernoulli distribution with parameter p.

Actually Lai and Robbins (1985) showed a slightly more general result that holds for all one-dimensional parametric distributions. However, the main extension has been performed in Burnetas and Katehakis (1996), in which the authors consider the case when the unknown distribution belongs to a known finite-dimensional parametric class of distributions $\mathcal{P}$. We introduce for that purpose the following quantity:

$$\mathcal{K}_{\inf}(\nu_a,\mu^\star) = \inf\big\{\mathcal{K}(\nu_a,\nu)\,;\, \nu\in\mathcal{P} \text{ with mean } \mu > \mu^\star\big\}\,,$$

with the convention that $\mathcal{K}_{\inf}(\nu_a,\mu^\star) = 0$ if the set $\{\nu\in\mathcal{P}; \mu > \mu^\star\}$ is empty.

Theorem 1.3 (*Distribution-dependent regret lower bound*) *Let us consider a set of probability distributions $\mathcal{P} \subset \mathcal{P}([0,1])$ and a forecaster consistent with $\mathcal{P}$. Then for any stochastic bandit with distributions in $\mathcal{P}$, the following holds true:*

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \geqslant \sum_{a:\Delta_a>0} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

The intuition for this improvement comes from the goal that we want to achieve in bandit problems; it is not detecting whether a distribution is optimal or not (for this goal, the relevant quantity would be $\mathcal{K}(\nu_a, \nu^\star)$), but rather achieving the optimal rate of reward $\mu^\star$, i.e. one needs to measure how close ν_a is to any distribution $\nu \in \mathcal{P}$ whose expectation is at least $\mu^\star$.

Let us also provide some additional intuition to explain why the right notion of closeness is a Kullback-Leibler-like divergence instead of other notions such as the Hellinger or Wasserstein distance. This comes from the proof where by definition of the regret it is enough to look at $\mathbb{E}(N_T(a))$ for each suboptimal arm a. For such an arm, we build a modified bandit problem where the distribution $\nu_a \in \mathcal{P}$ is replaced with a distribution $\nu_a^\star \in \mathcal{P}$ that has a mean bigger than $\mu^\star$ and satisfies $\nu_a \ll \nu_a^\star$, while other distributions remain the same. Now in this transported problem, a is the unique best arm, thus since we consider consistent algorithms, all the other arms are pulled $o(T^\beta)$ times for every β, and thus $N'_T(a)$, the number of pulls of arm a in the modified problem, is small with small probability w.r.t. $\nu_a^\star$. We then naturally control the number of pulls of arm a in the original problem by making use of a transport equation. The following one is used

$$\mathbb{E}_{\{X_i\}_{i\leqslant n}\sim\nu_a^\star} f(\{X_i\}_{i\leqslant n}) = \mathbb{E}_{\{X_i\}_{i\leqslant n}\sim\nu_a} f(\{X_i\}_{i\leqslant n}) \exp\Big(-n\widehat{\mathcal{K}}_n(\nu_a,\nu_a^\star)\Big)\,,$$

for some positive function f, where $\widehat{\mathcal{K}}_n(\nu_a,\nu_a^\star) \overset{\text{def}}{=} \dfrac{1}{n}\displaystyle\sum_{i=1}^n \log\frac{d\nu_a}{d\nu_a^\star}(X_i)$ is the empirical Kullback-Leibler divergence between the initial and the transported distribution and asymptotically converges to $\mathcal{K}(\nu_a,\nu_a^\star)$. Due to this transportation cost, we loose a little, but still it can be shown that $N_T(a) \leqslant \frac{\log(T)}{\mathcal{K}(\nu_a,\nu_a^\star)}$ asymptotically happens with vanishing probability. This gives an explanation for the final term, since $\nu_a^\star$ can then be chosen such that $\mathcal{K}(\nu_a,\nu_a^\star)$ is arbitrarily close to $\mathcal{K}_{\inf}(\nu_a,\mu^\star)$. We refer to Burnetas and Katehakis (1996) for specific details.

1.2 Historical algorithms

In this section, we now present some important historical algorithms that are designed in order to achieve performances that tempt to match the lower bounds on the expected-regret.

Asymptotically optimal strategies. In their seminal paper, Lai and Robbins (1985) provided an algorithm based on the Kullback-Leibler divergence that was proved to be asymptotically optimal for the case of some *one-dimensional* parametric distributions. This work

has been extended by Burnetas and Katehakis (1996) to an algorithm based on $\mathcal{K}_{\inf}$; this is still asymptotically optimal since the number of pulls of any sub-optimal arm a satisfies

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

The result holds for *finite-dimensional* parametric distributions under some assumptions, e.g., the distributions having a finite and known support or belonging to a set of Gaussian distributions with known variance. Recently Honda and Takemura (2010a) extended this asymptotic result to the case of arbitrary distributions $\mathcal{P}$ with support in $[0,1]$ and such that $\mu^\star < 1$; the key ingredient in this case is that $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$ is equal to $\mathcal{K}_{\min}(\nu_a, \mu^\star) \stackrel{\text{def}}{=} \inf_{\nu \in \mathcal{P}: E(\nu) \geqslant \mu^*} \mathcal{K}(\nu_a, \nu)$. Unfortunately, all these results provide *asymptotic* bounds only.

The Upper confidence bound (UCB) algorithm. The upper confidence bound (UCB) algorithm has been introduced by Auer et al. (2002), together with some important variants like UCB2, and enables to get *non-asymptotic* upper bounds. The main idea of the algorithm is detailed in Figure 1.2 and is to select at time t the arm a corresponding to the best empirical mean $\widehat{\mu}_{a,N_t(a)}$ penalized by some quantity $\delta_a(t)$. This quantity is typically a high probability upper confidence bound on the mean μ_a of arm a, thus justifying the name of the procedure, and is given by results from concentration inequalities.

Parameters: Penalty unctions $\delta_a : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ for all $a \in \mathcal{A}$

Initialization: Pull each arm of $\mathcal{A}$ once

For rounds $t+1$, where $t \geqslant |\mathcal{A}|$,

- compute for each arm $a \in \mathcal{A}$ the quantity

$$B_{t+1,N_t(a)}(a) = \widehat{\mu}_{a,N_t(a)} + \delta_a(t+1, N_t(a))$$

  where $$\widehat{\mu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} Y_s\,;$$

- in case of a tie, pick an arm with largest value of $\widehat{\mu}_{a,N_t(a)}$;
- pull any arm $A_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} B_{t+1,N_t(a)}(a)\,.$

Figure 1.2: The UCB-strategy.

The penalties δ_a of the initial algorithm UCB1 by Auer et al. (2002) are based on Hoeffding's inequality. A refined analysis have been performed in the same paper with the goal to get closer to the asymptotic distribution-dependent lower bound obtained by combination of

the result of Lai and Robbins (1985) together with Pinsker's inequality. Later in Audibert et al. (2009), Bernstein type inequalities with empirical variance estimate are used instead of Hoeffding's inequality. Finally, in Audibert and Bubeck (2010), another penalty is used with the goal to match the distribution-free lower bounds.

In order to give more intuition about these algorithms, we now provide a simple proof of the UCB algorithm, that corresponds to using the (non-optimal) penalty $\delta_a(t,s) = \sqrt{\frac{3\log(t)}{2s}}$. The other results follow by some refinements using different penalties and are mentioned after.

We first need to introduce some notations. We consider the filtration $(\mathcal{F}_t)$, where for all $t \geqslant 1$, the σ–algebra $\mathcal{F}_t$ is generated by $A_1, Y_1, \ldots, A_t, Y_t$. In particular, A_{t+1} and thus all $N_{t+1}(a)$ are $\mathcal{F}_t$–measurable. We denote by $\tau_{a,1}$ the deterministic round at which a was pulled for the first time and by $\tau_{a,2}$, $\tau_{a,3}$, ... the rounds $t \geqslant |\mathcal{A}| + 1$ at which a was then played; since for all $k \geqslant 2$,

$$\tau_{a,k} = \min\big\{t \geqslant |\mathcal{A}| + 1 : \quad N_t(a) = k\big\}\,,$$

we see that $\big\{\tau_{a,k} = t\big\}$ is $\mathcal{F}_{t-1}$–measurable. Therefore, for each $k \geqslant 1$, the random variable $\tau_{a,k}$ is a (predictable) stopping time. Hence, by Chow and Teicher 1988, Section 5.3, the random variables $\widetilde{X}_{a,k} = Y_{\tau_{a,k}}$, where $k = 1, 2, \ldots$ are independent and identically distributed according to ν_a.

Then we define the empirical mean of arm a as $\widetilde{\mu}_{a,s} = \dfrac{1}{s}\displaystyle\sum_{k=1}^{s} \widetilde{X}_{a,k}$

Theorem 1.4 (*Distribution-dependent regret bounds for UCB*) *In the stochastic multi-armed bandit game, the UCB algorithm strategy satisfies the following performance bound.*

$$R_T \leqslant \sum_{a;\Delta_a>0} \left[\frac{6}{\Delta_a}\log(T) + \Delta_a\frac{\pi^2}{3}\right]$$

Proof: Hoeffding's inequality state that for s i.i.d. random variables $X_i \in [0,1]$ with mean μ, we have

$$\mathbb{P}\big(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \geqslant \varepsilon\big) \leqslant e^{-2s\varepsilon^2} \quad \text{and} \quad \mathbb{P}\big(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \leqslant -\varepsilon\big) \leqslant e^{-2s\varepsilon^2}\,.$$

Thus, by application of this inequality to the random variables $\{\widetilde{X}_{a,k}\}_{k=1..s}$, we deduce that by definition of the penalization,

$$\mathbb{P}\big(\widetilde{\mu}_{a,s} + \delta_a(t,s) \leqslant \mu_a\big) \leqslant t^{-3} \quad \text{and} \quad \mathbb{P}\big(\widetilde{\mu}_{a,s} - \delta_a(t,s) \geqslant \mu_a\big) \leqslant t^{-3}\,.$$

Let us consider at time t the following event for a given arm a and $a^\star$,

$$\left\{\mu_{a^\star} - \delta_{a^\star}(t, N_{t-1}(a^\star)) \overset{(a)}{\leqslant} \widetilde{\mu}_{a^\star,N_{t-1}(a^\star)} \text{ and } \widetilde{\mu}_{a,N_{t-1}(a)} \overset{(b)}{\leqslant} \mu_a + \delta_a(t, N_{t-1}(a))\right\}. \tag{1.3}$$

Then if a is a suboptimal arm chosen at time t, then it means that $B_{t,N_{t-1}(a)}(a) \geqslant B_{t,N_{t-1}(a^\star)}(a^\star)$ for any optimal arm $a^\star$. Thus, we deduce from (1.3) that $\mu_a + 2\delta_a(t, N_{t-1}(a)) \geqslant \mu_{a^\star}$, i.e. $N_{t-1}(a) \leqslant \frac{6\log(t)}{\Delta_a^2}$.

Let us consider some integer u. We have:

$$\begin{aligned} N_T(a) &\leqslant u + \sum_{t=u+1}^{T} \mathbb{I}\{a_t = a \cap N_{t-1}(a) > u\} \\ &\leqslant u + \sum_{t=u+1}^{T} \mathbb{I}\{B_{t,N_{t-1}(a)}(a) \geqslant B_{t,N_{t-1}(a^\star)}(a^\star) \cap N_{t-1}(a) > u\}\,. \end{aligned}$$

Now, $\{B_{t,N_{t-1}(a)}(a) \geqslant B_{t,N_{t-1}(a^\star)}(a^\star)\}$ implies that either $N_{t-1}(a) \leqslant \frac{6\log(t)}{\Delta_a^2}$ or (1.3) is false. Thus, we set $u = \frac{6\log(t)}{\Delta_a^2}$ and deduce that either (a) or (b) is false. Since each of this event happens with probability less than t^{-3}, we deduce that by taking a union bound and the expectation in the previous expression, we get:

$$\begin{aligned} \mathbb{E}(N_T(a)) &\leqslant \frac{6\log(T)}{\Delta_a^2} + \sum_{t=u+1}^{T}\Big[\sum_{s=u+1}^{t} t^{-3} + \sum_{s=1}^{t} t^{-3}\Big] \\ &\leqslant \frac{6\log(T)}{\Delta_a^2} + \frac{\pi^2}{3}. \end{aligned}$$

□

We now detail the penalty functions δ_a that corresponds to the algorithm called UCB-α, UCB-V introduced in Audibert et al. (2009) and MOSS introduced in Audibert and Bubeck (2010). An upper bound on the pseudo-regret of these algorithms has been derived both for the distribution-free and distribution dependent case; they are gathered in the next two theorems. Note that if the MOSS algorithm nicely fills the gap with the distribution-free asymptotic bound, whereas UCB-α and UCB-V do not, there is still a gap between the proposed distribution-dependent bounds and the distribution-dependent asymptotic lower bounds. Also, in the following formulation, it has to know the time horizon T.

Theorem 1.5 (*Distribution-dependent regret bounds for UCB strategies*) *In the stochastic multi-armed bandit game, the UCB algorithm strategies satisfy the following performance bounds.*

Let $c_1(\alpha) = 1 + \frac{4}{\log(\alpha/2+1/2)}(\frac{\alpha+1}{\alpha-1})^2$. *Then provided* $\alpha > 1$, *UCB-*α *satisfies:*

$$R_T \leqslant \sum_{a;\Delta_a>0} \frac{2\alpha}{\Delta_a}\log(T) + \Delta_a c_1(\alpha)\,.$$

Let $c_2(\alpha) = 2 + \frac{12}{\log(\alpha+1)}(\frac{\alpha+1}{\alpha-1})^2$. *Then provided* $\alpha > 1$, *UCB-V satisfies:*

$$R_T \leqslant \sum_{a;\Delta_a>0} 8\alpha(\frac{\sigma^2(a)}{\Delta_a} + 2)\log(T) + \Delta_a c_2(\alpha)\,.$$

Upper Confidence Bound (UCB) algorithms
Parameter: $\alpha > 0$
For the arm a, define the quantity $\delta_a(t,0) = \infty$, and for $s \geqslant 1$:

$$\begin{aligned}
\text{UCB penalty:} \quad & \delta_a(t,s) = \sqrt{\tfrac{\alpha \log(t)}{2s}} \\
\text{UCB-V penalty:} \quad & \delta_a(t,s) = \sqrt{\tfrac{2\alpha V_s(a) \log(t)}{s}} + 3\alpha \tfrac{\log t}{s} \\
\text{MOSS penalty:} \quad & \delta_a(t,s) = \sqrt{\tfrac{\max(\log(\frac{T}{sA}),0)}{s}}
\end{aligned}$$

Figure 1.3: Upper Confidence Bound algorithms

Let $\Delta = \min_{a;\Delta_a>0} \Delta_a$. *Then MOSS satisfies:*

$$R_T \leqslant \frac{23A}{\Delta} \log\left(\max(\frac{110T\Delta^2}{K}, 10^4)\right).$$

Theorem 1.6 (*Distribution-free regret bounds for UCB*) *In the stochastic multi-armed bandit game, the UCB algorithm strategies satisfy the following performance bounds.*

Provided that $\alpha > 1$, *UCB-*α *satisfies:*

$$R_T \leqslant \sqrt{TA(2\alpha \log(T) + c_1(\alpha))}.$$

Provided that $\alpha > 1$, *UCB-V satisfies:*

$$R_T \leqslant \sqrt{TA(24\alpha \log(T) + c_2(\alpha))}.$$

Finally, MOSS satisfies:

$$R_T \leqslant 24\sqrt{TA}.$$

Kullback-Leibler Divergence based algorithms. In order to match the distribution-dependent lower bounds, we reanalyze in Chapter 2 the algorithm introduced by Lai and Robbins (1985) and later extended by Burnetas and Katehakis (1996) for which only asymptotic analysis was provided, and we thus provide a finite-time analysis for this algorithm.

The main idea is to focus on the empirical distribution itself instead of the empirical mean only, and use Kullback-Leibler divergence to compute an estimate of the distance to the best distribution.

This algorithm is detailed in Figure 1.4 and makes use of the empirical distribution of the first s rewards from ν_a, $\widehat{\nu}_{a,s} = \dfrac{1}{s}\displaystyle\sum_{k=1}^{s} \delta_{\widetilde{X}_{a,k}}$, instead of $\widehat{\mu}_{a,s}$.

Parameters: A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

Initialization: Pull each arm of $\mathcal{A}$ once

For rounds $t+1$, where $t \geqslant |\mathcal{A}|$,

– compute for each arm $a \in \mathcal{A}$ the quantity

$$B_{a,t}^{+} = \max\left\{q \in [0,1] : \quad N_t(a)\, \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,N_t(a)},\, q\big) \leqslant f(t)\right\},$$

where $$\widehat{\nu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} \delta_{Y_s}\,;$$

– in case of a tie, pick an arm with largest value of $\widehat{\mu}_{a,N_t(a)}$;

– pull any arm $A_{t+1} \in \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, B_{a,t}^{+}$.

Figure 1.4: The strategy $\mathcal{K}_{\inf}$.

The upper bound on the regret has been derived in the important case of distributions with a discrete support with finitely many points (see Chapter 2 for the precise statement). This result includes the special case of Bernoulli distributions and can also be extended to the case of one-dimensional exponential families. Note that the bound matches the corresponding asymptotic lower bound already derived in Burnetas and Katehakis (1996).

Theorem 1.7 (*Regret bound for the* $\mathcal{K}_{\inf}$*-strategy*) *Assume that $\nu^\star$ is finitely supported with support denoted by $\mathcal{S}^\star$ and expectation $\mu^\star \in (0,1)$ and that all all distributions ν_a are finitely supported. The expected regret of the $\mathcal{K}_{\inf}$-strategy, run with $f(t) = \log t$, satisfies for any $\varepsilon > 0$*

$$R_T \leqslant \sum_{a:\Delta_a > 0} \Delta_a \left[\frac{(1+\varepsilon)\,\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + c_a(\varepsilon) + \frac{1}{\varepsilon^2}\log\left(\frac{1}{1-\mu^* + \varepsilon}\right) \sum_{k=1}^{T} (k+1)^{|\mathcal{S}^\star|}\, e^{-k\varepsilon^2}\right],$$

where $c_a(\varepsilon)$ is an explicit constant.

2 Many extensions to the Bandit setting

The simple multi-armed bandit problem has been extended in many ways since its introduction by Robbins (1952). These extensions may concern different features of the bandit game, which sometimes involve specific analysis and theory. We now present these features, and then most of the known extensions to the stochastic multi-armed bandit setting according to those features.

For convenience, the general multi-armed bandit game is defined in figure 1.5. Let t refer to the time and $\mathcal{A}$ to the set of actions. We write $\mathcal{A}_t \subset \mathcal{A}$ the set of available actions at time t and r_t for the reward function provided by the environment at time t. At a high level, the game consists of choosing a sequence of actions a_t indexed by the time t in an online way, i.e. sequentially with time, according to observations received from the environment at each time step. The goal is to maximize some objective function that depends on the reward functions r_t sequentially defined by the environment.

General multi-armed bandit game
At time t, simultaneously:

- the learner chooses action $a_t \in \mathcal{A}_t \subset \mathcal{A}$.
- the environment defines a reward function $r_t : \mathcal{A} \to \mathbb{R}^+$.

then the learner outputs a_t, sees an information $i_t(r_t, a_t)$ from the environment, and receives (but may not see) a reward $r_t(a_t)$.

Figure 1.5: General multi-armed bandit game

Let us first present briefly the different features. We then detail each feature, providing motivations, algorithms and known results for most of them.

- The *power* of the environment: it refers to the way the environment is allowed to evolve. In the standard setting, the environment is considered to be a fix set $\mathcal{A}$ of sources of i.i.d. random variables. Now some natural extensions to the standard setting consider the case where the sources of i.i.d random variables are no longer fixed but may evolve with time. The way the source are allowed to evolve leads to different bandit settings (see the paragraphs about Adversarial bandits, Oblivious bandits, Restless bandits and Cooperative bandits below).

- The *information* received: it refers to the information received by the learner after playing arm a_t. In the standard setting, this information is $r_t(a_t)$ and is thus equaled to the received reward. In the case when we are only allowed to passively received information, natural extensions to this so-called Partial information setting include the settings of Full information, Side information and Trembling hand information. Now if we are able to actively grasp the information we want, other settings include the Label-efficient game, Budgeted learning and Slates bandit. Note again that this is not an assumption on the reward given to the learner, but only the information received about the reward function r_t.

- The *topological structure* of actions and rewards: it refers to the structure of the arms and of reward functions. In the standard setting, we consider a finite set of arms, and

uncorrelated reward functions. However in a lot of practical applications, the reward received after playing a specific arm gives some information on the value of the reward function on some other arms. This is particularly useful in order to deal with a large number of arms, including continuous set of arms, and different assumptions lead to very different settings. See paragraphs on Multi-armed bandit, Many-armed bandit, Convex bandit, Linear bandit, Lipschitz bandit and Topological bandit below.

- The *availability* of actions: it refers to the set of possible actions the learner is allowed to select an action from. In the standard setting, all actions are available at each time steps, thus we call this an awake bandit setting, however in many applications, it appears that not all actions are available at each time step but only a subset $\mathcal{A}_t$ of $\mathcal{A}$. This leads to important extensions on bandits. See Sleeping bandits, and Mortal bandits.

- The *regret* definition: it refers to the definition of the measure of performance, i.e. the regret. In the standard setting, we simply consider the cumulative pseudo-regret with respect to best constant strategy in hindsight. Many other assumptions may be considered, see paragraphs about External and Internal regret, Switching regret, Constrained regret, Simple, Maximum and Discounted regret below.

2.1 Power of the environment

If we consider the bandit problem from a game theoretical perspective, the hypothesis that the environment consists of A different fixed sources of i.i.d. random variables may appear as an arbitrary restriction on the power of the environment, that is seldom justified in practical situations such as games where the environment is a learner itself. This is the reason why people have considered different assumptions, and considered so-called *non-stochastic* bandits. However, allowing for a more powerful environment will generally make the standard notion of regret less justified, to the point that it may be preferable to consider the problem from a Reinforcement Learning (RL) perspective. We leave this philosophical point for Section 4 where we highlight some intrinsic limitations to the bandit approach, and we here focus on some variations around the standard stochastic bandit problem.

Adversarial bandits. The use of the word "adversarial" in the literature on bandit theory is a bit fuzzy. People generally refer to adversarial bandit to any case of bandit problem in the case when the rewards of the arms do not come from a fix distribution, but from a distribution evolving with time, and possibly adaptive to the learner. This means that in full generality the opponent is allowed to do virtually anything. Now in order to prevent confusion, we will talk of *oblivious* bandit when the opponent is not adaptive and of *adversarial* bandit when it is explicitly adaptive.

In order to prevent the opponent from learning our strategy, another class of algorithms has been introduced that differs from the deterministic UCB-based algorithms. Indeed being

deterministic against an adaptive opponent can lead to arbitrary bad performance. These algorithms are random algorithms in the sense that they define at each time step a probability distribution over the arms, then sample an arm according to this distribution and pull it.

Many definitions of regret can be considered for this setting. The most direct extension of the definition of the regret from the stochastic bandit setting is the cumulative regret with respect to the best constant strategy in hindsight, and is called the External regret.

Definition 1.2 (*External regret*) *The external regret for the adversarial bandit setting is defined by*

$$R_T = \max_{a \in \mathcal{A}} \mathbb{E}\Big(\sum_{t=1}^{T} r_t(a)\Big) - \mathbb{E}\Big(\sum_{t=1}^{T} r_t(a_t)\Big),$$

where the expectation is taken w.r.t. the possible randomization of the algorithm.

Note that in this definition, $r_t(a)$ is a shorthand for $r(h_{<t}, a)$, where $h_{<t}$ is the history of actions and rewards previously played up to time t. Thus we do not compare the cumulative reward obtained by the learner to the cumulative reward that would have been obtained by a player playing constantly the same action (and thus producing a different history $\tilde{h}_{<t}$). On the other hand, we keep the same history h_t in both terms of the difference, which is the meaning of the words "in hindsight".

The Exp3 algorithm (see Figure 1.6 in Section 3) introduced by Auer et al. (1995) manages to guarantee regret bounds of order $O(\sqrt{AT\log(A)})$ with respect to the best constant action in hindsight.

Note that it makes sense to define a regret reflecting the changes of the law of the rewards in the adversarial case. For instance, one may want to compete against the best sequence of actions instead of the best constant one. However, in this general case, it may be impossible (see Ryabko and Hutter (2008)) to learn the best sequence of actions to play. Now under certain conditions, we show in Chapter 4 how it is possible to achieve low regret against the best sequence of actions in hindsight.

Oblivious bandits. The setting of oblivious bandits considers the case when the rewards come from a distribution possibly evolving with time, but independently on the learner. Equivalently, one may consider that the rewards $r_t(a)$ for all time-step t and actions a are defined beforehand, i.e. before the beginning of the game. This means that the opponent is allowed to do anything but adapt to the learner.

In such a setting, the previous definition of (external) regret for the adversarial setting has clear interpretation, which gives an important motivation. Moreover, there are practical cases for which it is natural to consider that the environment is not adapting to our strategy, like in the stock market, provided that we trade sufficiently small amounts of money.

The Exp3 algorithm of Auer et al. (1995) also applies to this setting and give an upper bound of order $O(\sqrt{AT\log(A)})$. The main difference with the adversarial case is when

we want to consider a regret in high probability instead of a regret in expectation. The Exp3 algorithm has to be modified to guarantee high probability bounds in the adversarial case, by mixing the proposed probability distribution with a uniform distribution in order to guarantee sufficient exploration, while its genuine version is suitable for the oblivious case.

Restless bandits. The restless bandit is an attempt to compete with the changes in the law of the rewards in the oblivious case. Indeed one may consider that the probability laws on each arm evolves according to some internal (non observed) state structure. See Guha et al. (2007b) or Guha et al. (2007a) for further details.

One practical motivation for this setting is when we want to transmit some information on a canal, and that it is shared with other users. This has been studied extensively in Guha et al. (2007a) or Filippi (2010) for instance.

The regret that people consider for this setting is a strong notion. Actually, the comparison class considers the best sequence of actions and not just the best constant strategy, which is far more challenging. In Filippi (2010), it is proved that one may achieve a regret of order $O\big((\log T)^{1/3}T^{2/3}\big)$.

Cooperative bandits. Finally, an interesting feature that receives now increasing attention is that the opponent may not always be a foe. This appears for instance in so-called *cooperative games*, where the opponent and the player try to agree on some strategy. In such a setting, the opponent may try to help the learner get the minimal regret, but the opponent may not be fully cooperative due for instance to incomplete or noisy information received about the learner or the game. Thus it has a certain (unknown) degree of friendliness, and we want to be adaptive to this degree.

A practical motivation for this setting appears for instance in Brain Computer Interfaces (BCI), where the learner is an algorithm that observes the brain activity of a human (opponent). The human wants to cooperate with the learner in order to define meaningful symbols that he will use in order to pilot tasks using his mind only. Thus in this problem, the learner has to consider the environment as friendly, but since the human has no precise knowledge of the algorithm, he/she thus makes mistakes, even if he/she tries to be as friendly as possible.

This setting shares some links with the teaching theory Zilles et al. (2011) if we consider the viewpoint of the environment (teacher), and also with what is called learning with privileged information Pechyony et al. (2010), Vapnik and Vashist (2009) from the point of view of the learner. However, to the best of our knowledge, the results for this setting are quite limited.

2.2 Information received

So far, we have considered that the information given to the learner at time t is the reward of the selected action only, i.e. $r_t(a_t)$; this corresponds to the so-called *partial information*

setting. We present now other settings for which $i_t(r_t, a_t)$, the information received by the learner at time t after playing action a_t, may be different from $r_t(a_t)$.

2.2.1 Passive information retrieval

In the case of passive information retrieval, i.e. when the learner does not choose the information received, we can consider the following settings.

Full Information. The *full information* setting corresponds to the case when the information given to the learner at time t is the reward of all actions, i.e. $r_t(a)$ for all $a \in \mathcal{A}$, and not only the reward of the action chosen.

In the case of stochastic multi-armed bandit with finite arms with full information, there is no information retrieval challenge, thus for that reason the full information setting is generally considered in an *adversarial* setting, and/or when the set of arms is large, like for instance a convex subset of $\mathbb{R}^d$.

Since we receive more information than in the case of the partial information setting, it is not surprising that the upper and lower bounds on the regret for the partial and full information setting significantly differ. Indeed in the case of the adversarial multi-armed bandit setting with full information, it can be shown that the lower bound on the cumulative pseudo-regret is of order $O(\sqrt{T \log A})$ (see Cesa-Bianchi et al. (1993)), instead of $O(\sqrt{TA})$ in the Partial information setting (see Auer et al. (1995)); for that reason, the factor $\sqrt{\frac{A}{\log A}}$ is called the price for information in Dani et al. (2008a).

The weighted majority algorithm, introduced by Littlestone and Warmuth (1989), is the main tool to get a matching upper bound in the adversarial setting, and is better known as the Hedge algorithm, which is a variant introduced in Auer et al. (1995).

Side information. The side information setting (aka Contextual bandits) corresponds to the case when the information given at each round is the reward of the selected action a_t together with some other information $x_t \in \mathcal{X}$. We compare the forecaster to the best fixed hypothesis $h : \mathcal{X} \to \mathcal{A}$ in a class $\mathcal{H}$. See for instance chun Wang et al. (2005), Lu et al. (2010) or Lazaric and Munos (2009) for further details.

The motivation comes from the fact that in many practical problems, we have additional information on the information received. For instance, in the web-advertisement problem, where the goal is to select one add $a_t \in \mathcal{A}$ to display on a web-page, we know for instance what other information is displayed on the web-page, or some features about the human being surfing on this page thanks to cookies, or navigation history; one naturally wants to use this knowledge in order to adapt ones strategy to the human being.

Classical measures of the complexity include the Vapnik-Chervonenkis (VC) dimension and the Littlestone dimension of the class $\mathcal{H}$. In (Lazaric and Munos, 2009, p.9), there is a nice summary of the known results on the performance (regret) for this setting when considering different assumptions.

Trembling hand information. This setting corresponds to the case when the information given at each round is not the reward of the action a_t proposed by the learner, but the reward of a different action that is chosen by a third player. See Maillard and Munos (2011) and Chapter 4. In game theory, this is called a *trembling hand* effect.

A theoretical motivation for this apparently strange setting comes from the problem of learning using learner advices, instead of experts, for which a meta algorithm considers many learners in a partial information setting: at time t, each learner l proposes an action a_t^l, but after seeing the propositions the meta algorithm proposes a'_t instead, and get the reward $r_t(a'_t)$. Then the meta algorithm gives this information to each of the subordinate learners that are thus facing a bandit problem in the trembling hand information setting.

The corresponding notion of regret measures the performance of the algorithm that proposes at time t the distribution p_t but in the game where the action $a_t \sim q_t$ is played instead. Without any further assumption on the learners, the performance depends on the ratios between the probability $p_t(a'_t)$ of choosing arm a'_t defined by the learner, and the probability $q_t(a'_t)$ of choosing arm a'_t defined by the meta player. See Chapter 4 for further details as well as regret bounds.

2.2.2 Active information retrieval

In the case of active information retrieval, i.e. when we allow the learner to actively grasp information, we can consider at least the three following settings.

Label-efficient game. In this game, we consider that asking to see information is costly, which is motivated by practical implementations, and thus at each round, the forecaster chooses whether to see the reward(s) or not. We refer for instance to Cesa-Bianchi et al. (2005) or Allenberg et al. (2006) for important works in this domain.

In such a setting, we compare the cumulative reward of the algorithm to the cumulative reward of the best of A experts. The regret is of order $O(\sqrt{\frac{TA\log(A)}{\overline{m}}})$, where $\overline{m}$ is the average number of experts whose reward is revealed per round. Note that a value of $\overline{m}$ equal to A (resp. 1) corresponds to the full information (resp. partial information) setting. In the more general case when the learner can not choose to see the reward(s) more than m times over the T rounds i.e $\overline{m} = \frac{m}{T}$, the label-efficient regret is typically bounded by $O(T\sqrt{\frac{A\log(A)}{m}})$.

Budgeted learning. A closely related problem is called budgeted learning. In this setting, it is assumed that each information revealed is costly, with cost $c_t(a)$ at time t for arm a, and that additionally we have a total budget B that is limited.

One motivation comes from clinical trials, where each arm corresponds to one treatment, different treatments have different cost and we only have a limited budget to treat a patient. Thus, since some treatments may be very expensive, we do not want to use them too much unless they are very efficient.

However, it seems that there is little work on the problem for general costs functions, and most of the works only consider the case when $c_t(a) = 1$, see for instance Guha and Munagala (2007), while it would seem natural to consider different costs for different actions, or the more general case when c_t is only revealed at time t but otherwise arbitrary.

Slates bandit. In this setting, the learner is allowed to choose not only one but $p > 1$ arm per round. At each round the learner selects the set of p arms for which information is asked. This assumption modifies a lot the original bandit problem. It has been studied in Kale et al. (2010) and is a special case of the more general online linear optimization with bandit feedback problem (see Awerbuch and Kleinberg (2008) or Bartlett et al. (2008) for instance).

2.3 Topological structure

The information that the learner can infer about the all set of rewards $\{r_t(a)\}_{a,t}$ depends not only on the information effectively received by the learner, but also on what we call the topological structure of the problem, that basically states how far is some $r_s(a)$ from $r_t(b)$ for any s, t, a, b. Such assumptions on the problem may drastically change the achievable regret bounds.

Multi-armed bandit. The multi-armed bandit problem corresponds to the case when the set of arms $\mathcal{A}$ is finite, as considered from the beginning of this chapter. This situation naturally appears in a lot of problems, when we consider a production chain for instance. In this setting, there is generally no a priori relation between the rewards of arm a and rewards of arm b, except if we assume some correlation structure.

The multi-armed bandit problem with dependent arms has been addressed for instance in Pandey et al. (2007). They give a theoretical optimal policy and then provide a feasible algorithm that selects clusters of correlated arm before choosing one arm in this cluster.

Many-armed bandit. In this setting, we consider that the number or arms $\mathcal{A}$ is greater than the possible number of experiments and possibly infinite.

Typical applications for this setting include for instance labor markets, when a worker has many opportunities for jobs, or path planning in which the learner has to decide between a route that has proved to be efficient in the past (exploitation), or a known route that has not been explored many times (sampling), or a new route that has never been tried before (discovery).

A classical assumption in this context is that the distributions of rewards belong to a known parametric class, see Berry et al. (1997) and Poland (2008), and that the forecaster has a prior on the parameters of the arms. However in Wang et al. (2008), the authors make the weaker assumption that each arm has a probability ε^β of being ε-optimal, and then derive tight bounds depending on β.

Convex bandits. In this setting, we consider that the set of actions $\mathcal{A}$ is a (compact) convex subset of $\mathbb{R}^d$, and that the reward functions given by the opponent are convex.

In *online convex optimization* r_t is assumed to be a convex (Zinkevich (2003), Narayanan and Rakhlin (2010)) or σ-strongly convex (Hazan et al. (2006)) function of a. The resulting upper bounds are of order $C\sqrt{T}$ and $C^2\sigma^{-1}\ln(T)$, where C is a bound on the gradient of the functions, which implicitly depends on the space dimension. Other extensions have been considered in Bartlett et al. (2007), Shalev-Shwartz (2007), Flaxman et al. (2005) and a minimax lower bound analysis in the full information case is provided in Abernethy et al. (2008a). These results hold in a bandit information setting where either the value or the gradient of the function is revealed.

Linear bandits. In this setting, we consider the case when the functions r_t are linear,i.e. $r_t(a) = L_t \cdot a_t$ for some $L_t \in \mathbb{R}^d$. This setting has recently received increasing interest due to the link with practical applications, see e.g. Dani et al. (2008a), Abernethy et al. (2008b), Cesa-Bianchi and Lugosi (2009), Kakade et al. (2008) in the adversarial case, Auer (2003), Dani et al. (2008b) in the stochastic case, and also Auer (2003), Awerbuch and Kleinberg (2008), Abernethy et al. (2008b), Dani et al. (2008b), Rusmevichientong and Tsitsiklis (2010).

The resulting upper and lower bounds on the regret are, up to logarithmic factors, of order $\sqrt{dT}$ in the case of full information, $d^{3/2}\sqrt{T}$ in the case of bandit information Abernethy et al. (2008b), and $d\sqrt{T}$ in the good cases, seeDani et al. (2008a) for details.

Lipschitz bandits. In this setting, we consider the weaker assumption that the reward function are Lipschitz functions with constant λ. This has been considered in Kleinberg et al. (2008), Bubeck et al. (2008) for the stochastic bandit setting, and in Maillard and Munos (2010b) for the adversarial setting. See also Chapter 3.

One motivation for this setting is that the assumption that the reward functions are convex is often too strong in practice. The weaker Lipschitz assumption enables to apply the bandit setting to non-convex problems, like for instance online regression with non convex regression functions (e.g. feed-forward neuron-networks).

The regret bound in the partial information setting is $\tilde{O}(\lambda^{\frac{d}{d+2}}T^{\frac{d+1}{d+2}})$ where d is the pseudo-dimension of problem, defined for instance in Bubeck et al. (2008). In the full information adversarial bandit setting one can derive a bound of order $\sqrt{dT\ln(\lambda dT)}$ that is achieved by some variant of an exponentially weighted forecaster, see details in Maillard and Munos (2010b) and chapter 3.

Metric bandits. Finally in the stochastic setting, we can even relax the Lipschitz assumption. In Bubeck et al. (2008), it is only assumed that the set of arms $\mathcal{A}$ is a metric space and that the mean-payoff function that maps each arm to the average payoff one receives by pulling this arm is weakly Lipschitz around its maximum (see Bubeck et al. (2008) for precise definition). The regret bounds in this case scales with the near-optimality dimension

d' of the mean-payoff function as $\tilde{O}(T^{\frac{d'+1}{d'+2}})$. Note that d' is typically much smaller than the ambient dimension d.

2.4 Availability of actions

So far, we have assumed that all actions $a \in \mathcal{A}$ are available at each time step. In many practical situations it appears that only a subset $\mathcal{A}_t \subset \mathcal{A}$ is available at time t. This may happen for example in the case when the arms correspond to experts and that one expert is not available at some point. This modifies the standard definition of the regret as well as the standard algorithms and analysis.

Sleeping bandits. The sleeping bandit setting considers the case when an expert a is awake from time to time, say when $t \in I(a) \subset \{1, \dots, T\}$. See Kleinberg et al. (2008) and Kanade et al. (2009). Three main situations are considered, Deterministic availability, Stochastic availability and Adversarial availability, according to whether $I(a)$ is defined deterministically, chosen according to some distribution, or depending on the learner.

In Kanade et al. (2009), the authors propose a (tractable) algorithm that achieve a regret bounded by $O((TA)^{4/5}\log(T))$ in the case of stochastic availability.

One may note that this problem can be recast in the setting of classical (all awake) bandits by considering that one arm in the modified problem is one permutation of the arms of the initial problem, which leads to a standard bandit problem with $\mathcal{A}!$ all awake many arms. Thus there exist algorithms that achieve $\sqrt{T\log(A!)}$ regret bounds, but unfortunately, due to the transformation considered, the final algorithm is not tractable. Thus the major issue in this setting is to address the tractability problem. Actually, it is still an open question to know if one can design *tractable* algorithms with performance of order $\tilde{O}(\sqrt{T})$.

Mortal bandits. A special case of sleeping bandits is when we assume that each expert can only be awake for a total amount of time that is bounded, and then dies, see Chakrabarti et al. (2008). This case is for instance motivated by web-advertisement, where the goal is to display one add amongst many possible with the goal to have the displayed add clicked and where in practice, a publicist will sign a contract selling some amount of $d(a)$ displays for the add a. Thus after the add a has been displayed $d(a)$-many times, it can not be displayed more, which motivates this setting. Due to the mortality of arms, deciding when to display an add or not becomes a quite complicated task.

2.5 Other regret definitions

Finally the definition of the regret may also vary.

External and Internal regret. The notion of *internal* regret (see Foster and Vohra (1996)) compares the loss of an online algorithm to the loss of a modified algorithm that

consistently replaces one action by another, and has been also considered in many works Hart and Mas-Colell (2000), Stoltz (2005), Cesa-Bianchi and Lugosi (2003), Foster and Vohra (1999). In Blum and Mansour (2005), the authors propose a way to convert any external regret minimization algorithm into an algorithm minimizing an extended notion of internal regret, using the so-called modifications rules that are functions $h, a \to b$, where h is a possible history, and a and b are actions.

Best switching strategy regret. In Auer et al. (2003), the authors extend the class of comparison strategies to piecewise constant strategies with at most S switches. The corresponding Exp3S (aka ShiftBand) algorithm achieves a regret of order $\sqrt{TSA\log(T^3A)}$, provided that T is large enough.

Best constrained strategy regret. In Maillard and Munos (2011), we consider general classes of strategies that are defined as mappings from equivalence classes of histories to actions, based on some equivalence relation φ (see Chapter 4). With this notion, we can define a corresponding regret with respect to the best such strategy. For instance, using the trivial equivalence relation such that all histories belong to the same class, we recover the standard notion of regret, but such a definition captures more expressive notions of regret, from classical regret to the regret w.r.t. the best (possibly switching at each time step) strategy in hindsight.

A lower bound on the regret is of order $\sqrt{TAC}$ where C is the complexity of the opponent, defined as the number of classes of equivalence of histories he uses.

We develop in Chapter 4 efficient algorithms that match this lower bound in the case when the model $\varphi^\star$ of the opponent is known and with bound of order $(TA)^{2/3}C^{1/3}\log(|\Phi|)^{1/2}$ in the case it is unknown and we use a finite set Φ of possible models.

Simple, Maximum and Discounted regret. In some problems, we may not be interested in the cumulative regret, but in some other notion. For instance, in the setting of online learning, we only care about the final proposed action a_T at time T, whatever the actions previously taken. The corresponding notion of regret measures the performance $\mathbb{E}(r(a_T)-r(a^\star))$ at final time T, as opposed to the cumulative sum over all time steps, and is also known as the simple regret, see Bubeck et al. (2009).

More generally, one may want to consider a discounted cumulative regret, with discount factor $\gamma \in [0,1]$, where the cumulative performance is now

$$\sum_{t=1}^{T} \gamma^{T-t} r_t(a_t),$$

which enables to both recover the cumulative regret (with $\gamma = 1$) and the simple regret (with $\gamma = 0$). One may also want to consider the performance criterion to be the maximal instantaneous regret along all time steps, i.e. $\max_{1\leqslant t\leqslant T}\mathbb{E}(r(a_t) - r(a^\star))$.

Beyond regret. Finally, let us mention that in Rakhlin et al. (2010), the authors study a very general notion of regret that generalizes all previously defined notions. Their goal is to study the notion of regret from a minimax point of view and not from an algorithmic point of view. They define a nice notion of sequential complexity and show how a control of this complexity may lead to a tight control on the corresponding notion of regret. They derive minimax optimal bounds accordingly, although these derivations, as pointed out in the article, do not lead to numerically efficient algorithms. Despite this numerical problem, I believe this opens a very nice area of research for the future of bandit theory.

Conclusion. The previous classification of the bandit literature is already big, but unfortunately can not handle the huge variety of problem variations around the standard bandit problem - there are actually several thousands publications about bandits (!). For instance, I did not talk about Bayesian approaches to the bandit problem, with probabilistic procedures such that the now famous Gittins' indices, see Whittle (1980), Gittins et al. (1989).

3 Exponentially-weighted decision-makers

In this section, we now focus on a class of algorithms known as exponentially-weighted decision-makers (aka exponentially-weighted forecasters). These algorithms was initially designed to handle the setting of adversarial bandits and thus are robust to an arbitrary bad opponent in some sense. We here explain this phenomenon first in the partial information setting, and then provide exact computation of the performance bound in the case of full information. Finally we show that one can also design a similar algorithm that achieves a good performance within the setting of stochastic bandits, which was not expected according to popular belief.

First of all, we introduce the notion of exponential families that plays an important role in the interpretation of exponentially-weighted forecasters.

3.1 Exponential families

In this section, we write $\mathcal{P}(\mathcal{A})$ the set of probability distributions on the set $\mathcal{A}$.

Definition 1.3 (*Exponential families*) *The exponential family generated by the set of functions $(F_k)_{k\leqslant K}$ and the reference measure ν_0 on the set $\mathcal{A}$ is*

$$\mathcal{E}((F_k)_{k\leqslant K};\nu_0)=\left\{\nu_\theta\in\mathcal{P}(\mathcal{A})\,;\,\nu_\theta(a)=\exp\Big(\sum_{k=1}^K\theta_kF_k(a)-z(\theta)\Big)\,,\,\theta\in\mathbb{R}^K\right\}\,,$$

where $z(\theta)\stackrel{\text{def}}{=}\log\int_{\mathcal{A}}\exp(\sum_{k=1}^K\theta_kF_k(a))\nu_0(da)$ is the normalization function of the exponential family. The vector θ is called the vector of canonical parameters.

An interesting property of exponential families is the following straightforward identity:

$$\mathcal{K}(\nu_{\theta_1}, \nu_{\theta_2}) = \langle \theta_1 - \theta_2, \mathbb{E}_{a\sim\nu_{\theta_1}}(F(a))\rangle - z(\theta_1) + z(\theta_2)\,, \tag{1.4}$$

where $F(a) \in \mathbb{R}^K$ is the vector with kth component $F_k(a)$. In particular, the vector $\mathbb{E}_{a\sim\nu_{\theta_1}}(F(a))$ is called the vector of dual (or expectation) parameters. It is equal to the vector $z'(\theta_1)$.

Let us remind also that in the general case, i.e. not only for the special case of exponential families, the Kullback-Leibler divergence can always be written in its variational form as

$$\mathcal{K}(\nu_1, \nu_2) = \sup\Big\{ \int_0^1 \varphi d\nu_1 - \log \int e^{\varphi} d\nu_2 \,;\ \varphi \in \mathcal{C}_b([0,1]) \Big\}\,,$$

where $\mathcal{C}_b([0,1])$ is the set of continuous and bounded functions on $[0,1]$.

3.2 Adversarial rewards with partial information

In the adversarial setting with partial information, at each round t the learner pulls one arm a_t and only gets to see the reward of the chosen arm. The following algorithm described in Figure 1.6 enables to guarantee high performance bounds.

Parameters: Increasing sequence $\{\eta_t\}_{1\leqslant t\leqslant T}$ For all $t = 1..T$,

- Define the losses
$$\tilde{l}_s(a) \stackrel{\text{def}}{=} \frac{1 - r_s(a)}{p_s(a)} \mathbb{I}[a_s = a]\,.$$
- Define the distribution
$$p_{t,\eta_t}(a) \stackrel{\text{def}}{=} \frac{\exp(-\eta_t \sum_{s=1}^{t-1} \tilde{l}_s(a))}{\sum_b \exp(-\eta_t \sum_{s=1}^{t-1} \tilde{l}_s(b))}\,.$$
- Select the arm $a_t \sim p_{t,\eta_t}$.

Figure 1.6: The Exp3 algorithm

An interesting point of view is to consider that the probability distributions p_{t,η_t} defined at each round t by the Exp3 algorithm belong to the same (random) exponential family $\mathcal{E}((\tilde{l}_s)_{1\leqslant s\leqslant T}, \nu_0)$; we introduce the parameter

$$\theta_t = \theta_t(\eta_t) \stackrel{\text{def}}{=} (-\eta_t, \ldots, -\eta_t, 0, \ldots, 0) \in \mathbb{R}^T\,,$$

with $t-1$ non zero values, and we get the rewriting

$$p_{t,\eta_t}(a) = p_{\theta_t}(a) = \exp\big(\sum_{s=1}^{T} \theta_t \tilde{l}_s(a) - z(\theta_t))\big)\,,$$

where $z(\theta) = \log(\sum_{a\in\mathcal{A}} \exp\big(\sum_{s=1}^{T} \theta \tilde{l}_s(a)\big))$ is the normalization function of the exponential family. Note that since $\theta_1 = 0 \in \mathbb{R}^T$, p_{θ_1} is the uniform distribution; we rewrite it U for convenience. Now by direct application of equation (1.4) we have the property that:

$$\frac{1}{\eta_t}\mathcal{K}(p_{\theta_1}, p_{\theta_t}) + \mathbb{E}_{a\sim U}(\sum_{s=1}^{t-1} \tilde{l}_s(a)) = \frac{z(0) - z(\theta_t)}{-\eta_t}\,,$$

which justifies the introduction of the function $\Phi_t(\eta) \stackrel{\text{def}}{=} \frac{z(0)-z(\theta_t(\eta))}{-\eta}$.

Before deriving the upper bound on the regret of the Exp3 algorithm, we need the following second order Taylor approximation, which is the the key stone of the proof of the next theorem, see also insights from Section 2 of chapter 5 about PAC-analysis.

Lemma 1.1 *Let X be a positive random variable, and $M_X(-\eta) = \mathbb{E}_X e^{-\eta X}$ be its moment generating function. Then for all $\eta \geqslant 0$:*

$$\log(M_X(-\eta)) \leqslant -\eta\mathbb{E}(X) + \frac{\eta^2}{2}\mathbb{E}(X^2)\,.$$

Proof: This directly follows by definition of M_X, together with the fact that for all $x \geqslant 0$, then $e^{-x} \leqslant 1 - x + x^2/2$, and $\log(x) \leqslant x - 1$. □

Theorem 1.8 (*Regret bound for Exp3*) *Provided that for all t, $\eta_t \geqslant \eta_{t+1}$, then the following holds true:*

$$R_T \leqslant \frac{\log(A)}{\eta_T} + \frac{A}{2}\sum_{t=1}^{T} \eta_t.$$

Proof: The proof is divided in five steps.

Step 1. Rewrite the regret term to make appear the decision probability p_t. Since $\mathbb{E}_{a\sim p_{t,\eta_t}} \tilde{l}_t(a) = 1 - r_t(a_t)$ and $\mathbb{E}_{a_t\sim p_{t,\eta_t}} \tilde{l}_t(a) = 1 - r_t(a)$, we deduce that for all $a \in \mathcal{A}$:

$$\sum_{t=1}^{T} r_t(a) - r_t(a_t) = \sum_{t=1}^{T} \mathbb{E}_{a\sim p_{t,\eta_t}} \tilde{l}_t(a) - \sum_{t=1}^{T} \mathbb{E}_{a_t\sim p_{t,\eta_t}} \tilde{l}_t(a)\,.$$

Step 2. Since we are interested in the random variable $X = \tilde{l}_t(a)$, where a is distributed according to p_{t,η_t}, conditionally on the other random variables, we introduce $M(\eta; \tilde{l}_t)$ the

moment-generating function of X, i.e. $M(\eta;\tilde{l}_t) = \mathbb{E}_X(e^{\eta X}) = \mathbb{E}_{a\sim p_{t,\eta_t}}(e^{\eta\tilde{l}_t(a)})$. Since X is a positive random variable, Lemma 1.1 applies; we deduce that

$$\mathbb{E}_{a\sim p_{t,\eta_t}}(\tilde{l}_t(a)) \leqslant -\frac{1}{\eta_t}\log(M(-\eta_t;\tilde{l}_t)) + \frac{\eta_t}{2}\mathbb{E}_{a\sim p_{t,\eta_t}}(\tilde{l}_t(a)^2)\,.$$

Step 3. In order to compute the term $\log(M(-\eta_t;\tilde{l}_t))$, we use the definition of the probability distribution p_{t,η_t} in terms of exponential family; we have

$$\log(M(-\eta_t;\tilde{l}_t)) = \log\left(\frac{\sum_{a\in\mathcal{A}} e^{-\sum_{s=1}^{t}\eta_t\tilde{l}_s(a)}}{\sum_{a\in\mathcal{A}} e^{-\sum_{s=1}^{t-1}\eta_t\tilde{l}_s(a)}}\right) = z(\theta_t(\eta_t)) - z(\theta_t(\eta_{t-1}))\,.$$

Thus, by definition of the function Φ_t, we deduce that:

$$\sum_{t=1}^{T}\mathbb{E}_{a\sim p_{t,\eta_t}}(\tilde{l}_t(a)) \leqslant \sum_{t=1}^{T}\Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) + \sum_{t=1}^{T}\frac{\eta_t}{2}\mathbb{E}_{a\sim p_{t,\eta_t}}(\tilde{l}_t(a)^2)\,. \tag{1.5}$$

Step 4. We bound each term.

First, since the reward function is bounded by 1 we have:

$$\mathbb{E}_{a\sim p_{t,\eta_t}}(\tilde{l}_t(a)^2) \leqslant \frac{1}{p_t(a_t)}.$$

Then, since $\Phi_0(\eta_1) = 0$, we can rewrite the first sum on the left hand of equation (1.5) as

$$\sum_{t=1}^{T}\Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) = -\Phi_T(\eta_T) + \sum_{t=1}^{T-1}\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)\,.$$

Now using the fact that the sum of positive terms is bigger than any of its term, $-\Phi_T(\eta_T)$ is bounded for all $a\in\mathcal{A}$ by:

$$\begin{aligned} -\Phi_T(\eta_T) &= \frac{\log(A)}{\eta_T} - \frac{1}{\eta_T}\log\big(\sum_b \exp(-\sum_{t=1}^{T}\eta_T\tilde{l}_t(b))\big) \\ &\leqslant \frac{\log(A)}{\eta_T} + \sum_{t=1}^{T}\tilde{l}_t(a)\,. \end{aligned}$$

Now in order to take care of the remaining term, we look at $\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)$. Since the sequence $(\eta_t)_t$ is by assumption decreasing with t, it suffices to show that for all t, the function Φ_t is non decreasing ; by definition of Φ_t and then by application of equation (1.4) we have

$$\begin{aligned} \Phi_t'(\eta) &= \frac{1}{\eta^2}[\,\langle z'(\theta_t(\eta),\theta_t'(\eta)\rangle\,\eta + z(0) - z(\theta_t(\eta)] \\ &= \frac{1}{\eta^2}[-\mathbb{E}_{a\sim p_{t,\eta}}\Big(\sum_{s=1}^{t-1}\tilde{l}_s(a)\Big)\eta + z(0) - z(\theta_t(\eta))] \\ &= \frac{1}{\eta^2}\mathcal{K}(p_{t,\eta},U) \geqslant 0\,. \end{aligned}$$

Step 5. We conclude by taking the expectation over all the remaining random variables, noticing that $\mathbb{E}(\frac{1}{p_t(a_t)}) = A$; we get:

$$R_T = \mathbb{E}(\sum_{t=1}^{T} \mathbb{E}_{a\sim p_t}(\tilde{l}_t(a)) - \sum_{t=1}^{T} \mathbb{E}_{a_t\sim p_t}(\tilde{l}_t(a))) \leqslant \frac{\log(A)}{\eta_T} + \frac{A}{2}\sum_{t=1}^{T}\eta_t\,.$$

□

Of course there are many extensions to this simple algorithm and analysis. For instance one may get regret bounds not only in expectation but also in high probability, see the algorithm Exp3.P by Auer et al. (2003), or may also want to replace Hoeffding's Lemma with the following Bernstein version,

$$\log(M_X(-\eta)) \leqslant (e^{-\eta} - 1)\mathbb{E}(X)\,,$$

and get the corresponding regret bound that maybe useful in the case when one the optimal arm has rewards very closed to 1. Following Peter Bartlett, we state it here for the simple case when $\eta_t = \eta$ for all t:

$$R_T \leqslant \frac{\log(A)}{1-e^{-\eta}} + (\frac{\eta}{1-e^{-\eta}} - 1)(T - \max_{a\in\mathcal{A}}\sum_{t=1}^{T} r_t(a))\,.$$

3.3 Adversarial rewards with full information

In the adversarial setting with full information, we now assume that there is a set of (unknown) reward functions $(r_t)_{1\leqslant t\leqslant T}$ defined before the game. At the end of each round t, we observe the full function r_t. The following Hedge algorithm enables to achieve high performance bounds in the worst case scenario. Note that here the set $\mathcal{A}$ does not need to be finite. In the case $\mathcal{A}$ is finite, the reference measure ν_0 is simply the counting measure, while in the general case it is chosen so that $\nu_0(\mathcal{A}) < \infty$. However in the later case, the following algorithm is only theoretical for it may not be possible to sample exactly from the distributions proposed p_t. This is discussed for instance in Narayanan and Rakhlin (2010) and in Maillard and Munos (2010b).

A regret analysis similar to the partial information setting can be done for this algorithm, and using standard tools from convex analysis, we can extend both algorithm and analysis to provide very powerful results for online prediction that we do not report here (see for instance Bartlett et al. (2007), Stoltz (2005)). Note, however, that the optimal value for the parameter $\eta = \eta_t$ is still an opened question, as explained for instance in chapter 3 of Stoltz (2011), where it is shown that a data-dependent value for η exhibits much better behavior in practice than the standard data-independent optimal value (yet with no theoretical analysis).

The following theorem shows another nice property of this algorithm. and was mentioned in Narayanan and Rakhlin (2010). Here, we derive this result using exponential families.

Parameter: $\eta > 0$, reference measure ν_0
For all $t = 1..T$,

- Define the distribution

$$p_t(a) = \frac{\exp(-\eta \sum_{s=1}^{t-1} r_s(a))}{\int_{\mathcal{A}} \exp(-\eta \sum_{s=1}^{t-1} r_s(b))\nu_0(db)}.$$

- Select the arm $a_t \sim p_{t,\eta}$.

Figure 1.7: The Hedge algorithm

Theorem 1.9 (*Full information performance bound for Hedge*) *Let A be a random variable with distribution q, and consider the algorithm Hedge. Then we have the following equality,*

$$\mathbb{E}[\sum_{t=1}^{T} r_t(a_t) - \sum_{t=1}^{T} r_t(A)] = \eta^{-1}(\mathcal{K}(q, p_1) - \mathcal{K}(q, p_{T+1})) + \eta^{-1} \sum_{t=1}^{T} \mathcal{K}(p_t, p_{t+1}).$$

Proof: We write $p_t = p_{\theta_t(\eta)}$ and thus see p_t as a member of the exponential family generated by $(r_t)_{t \leqslant T}$, with parameter

$$\theta_t(\eta) = (-\eta, \dots, -\eta, 0, \dots, 0) \in \mathbb{R}^T .$$

By direct application of equation (1.4), we have that

$$\mathcal{K}(p_{\theta_t(\eta)}, p_{\theta_{t+1}(\eta)}) = \eta \mathbb{E}_{a \sim p_{\theta_t(\eta)}}(r_t(a)) - z(\theta_t(\eta)) + z(\theta_{t+1}(\eta)) ,$$

thus we deduce that

$$\sum_{t=1}^{T} \mathbb{E}_{a \sim p_{\theta_t(\eta)}}(r_t(a)) = \frac{1}{\eta} \sum_{t=1}^{T} \mathcal{K}(p_{\theta_t(\eta)}, p_{\theta_{t+1}(\eta)}) + \frac{1}{\eta}[z(\theta_1(\eta)) - z(\theta_{T+1}(\eta))] .$$

Now by definition of the normalization function z, we have for all $a \in \mathcal{A}$,

$$z(\theta_1(\eta)) - z(\theta_{T+1}(\eta)) = \log \left(\frac{p_{\theta_{T+1}(\eta)}(a)}{p_{\theta_1(\eta)}(a)} \right) + \eta \sum_{s=1}^{T} r_s(a) .$$

Thus, by taking the expectation over q and reorganizing the terms, we get:

$$\mathbb{E}_{a \sim q}\Big(\sum_{s=1}^{T} r_s(a) \Big) = \frac{1}{\eta}[\mathcal{K}(q, p_{\theta_1(\eta)}) - \mathcal{K}(q, p_{\theta_{T+1}(\eta)})] + \frac{1}{\eta}[z(\theta_1(\eta)) - z(\theta_{T+1}(\eta))] ,$$

which concludes the proof. □

This Theorem is interesting especially since it provides an equality and not only an inequality on the performance of the algorithm, also due to presence of the target distribution q. For instance, the external regret corresponds to the case when q is a Dirac distribution. In a more general case, this result is the basis used in Narayanan and Rakhlin (2010) in order to prove a regret bound for this algorithm when the reward functions are assumed to be linear. See chapter 3 for an analysis of this algorithm in the case when the reward functions are assumed to be only Lipschitz.

3.4 Stochastic rewards with partial information

The exponentially weighted algorithms have been introduced to handle the adversarial bandit setting, which fundamentally differs from the stochastic bandit setting, and thus it is generally thought that they are not suitable for the stochastic setting. However, it appears that, interestingly enough, one can actually define such an algorithm using exponential weights to get high performance bounds in the stochastic setting as well. This is what we show in the following analysis, where we introduce the algorithm that we call the exponentially weighted stochastic algorithm, or simply EwS.

For all $t = 1..T$,

- Define the empirical gaps
$$\widehat{\Delta}_{a,t} \stackrel{\text{def}}{=} \max_{a' \in \mathcal{A}} \widehat{\mu}_{a',N_t(a')} - \widehat{\mu}_{a,N_t(a)} \,.$$
- Define the distribution
$$p_t(a) \stackrel{\text{def}}{=} \frac{\exp(-2N_{t-1}(a)\widehat{\Delta}^2_{a,t-1})}{\sum_{b\in\mathcal{A}} \exp(-2N_{t-1}(b)\widehat{\Delta}^2_{b,t-1})} \,.$$
- Select the arm $a_t \sim p_t$.

Figure 1.8: The exponentially weighted stochastic (EwS) algorithm

Theorem 1.10 (*Regret bound for EwS*) *For all $c_a > 0$, the expected number of times the Exponentially-weighted stochastic algorithm (EwS) pulls arm a satisfies*

$$\mathbb{E}\big[N_T(a)\big] \leqslant \frac{(1+c_a)^2 \log(T)}{2\Delta_a^2} + \frac{1}{|\mathcal{A}|} + 1 + C(c_a) \,,$$

where

$$C(c_a) \leqslant \frac{4(1+c_a)^2}{c_a^2\Delta_a^2}\big[1+\frac{32|\mathcal{A}|(1+c_a)^2}{c_a^2\Delta_a^2}\big]\,.$$

Note that the constant $C(c_a)$ in the above Theorem is not optimized and a more careful analysis may lead to better constants. The interesting part is the leading term that depends on $\log(T)$. Indeed, Pinsker's inequality entails that the lower bound $\mathcal{K}(\nu_a,\nu^\star) \geqslant 2\Delta_a^2$ is an optimal (first-order) approximation and thus the asymptotic behavior $\dfrac{\log(T)}{2\Delta_a^2}$ is the right dependency. Note also that by optimizing over c_a, we get exactly the leading term $\frac{\log(T)}{2\Delta_a^2}$ (with constant 1) and a second order term $O(\log(T)^{2/3})$.

Proof: **Step 1.** First, using the definition of the initialization step, for all $u>0$ we have

$$\begin{aligned}\mathbb{E}\big[N_T(a)\big] &\leqslant u+\mathbb{E}\Big(\sum_{t=1}^{T}\mathbb{I}_{\left\{A_t=a \text{ and } N_{t-1}(a)>u\right\}}\Big)\\ &\leqslant u+\frac{1}{|\mathcal{A}|}+\mathbb{E}\Big(\sum_{t=1}^{T-1}\mathbb{I}_{\left\{A_{t+1}=a \text{ and } N_t(a)>u\right\}}\Big)\,.\end{aligned}$$

Then, we also have the following inclusion of events

$$\Big\{\mu^\star-\varepsilon\leqslant\max_{a\in\mathcal{A}}\widehat{\mu}_{a,N_t(a)}\Big\}\cap\Big\{\widehat{\Delta}_{a,t}\leqslant\frac{\Delta_a}{1+c}\Big\}\subseteq\Big\{\widehat{\mu}_{a,N_t(a)}-\mu_a\geqslant\frac{c\Delta_a}{1+c}-\varepsilon\Big\}\,.$$

Indeed, on the first event, we have that $\dfrac{\Delta_a}{1+c}\geqslant\Delta_a-\varepsilon+\mu_a-\widehat{\mu}_{a,N_t(a)}$. Therefore, we deduce that

$$\begin{aligned}\mathbb{E}\big[N_T(a)\big] &\leqslant u+\frac{1}{|\mathcal{A}|}+\sum_{t=1}^{T-1}\mathbb{P}\Big\{\mu^\star-\varepsilon>\max_{a\in\mathcal{A}}\widehat{\mu}_{a,N_t(a)} \text{ and } N_t(a)>u \text{ and } A_{t+1}=a\Big\}\\ &\quad+\sum_{t=1}^{T-1}\mathbb{P}\Big\{\widehat{\mu}_{a,N_t(a)}-\mu_a\geqslant\frac{c\Delta_a}{1+c}-\varepsilon \text{ and } A_{t+1}=a \text{ and } N_t(a)>u\Big\}\\ &\quad+\sum_{t=1}^{T-1}\mathbb{P}\Big\{\widehat{\Delta}_{a,t}>\frac{\Delta_a}{1+c} \text{ and } A_{t+1}=a \text{ and } N_t(a)>u\Big\}\,. \qquad (1.6)\end{aligned}$$

Step 2. The third sum in (1.6) is handled by definition of the algorithm, since by definition of p_{t+1}, we have that

$$\mathbb{P}\Big\{A_{t+1}=a\,|\,\widehat{\Delta}_{a,t}>\frac{\Delta_a}{1+c} \text{ and } N_t(a)>u\Big\}\leqslant\exp(-2u\big(\frac{\Delta_a}{1+c}\big)^2)\,;$$

now by choosing $u\overset{\text{def}}{=}\frac{(1+c)^2\log(T)}{2\Delta_a^2}$ we deduce that

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{P}\Big\{\widehat{\Delta}_{a,t}>\frac{\Delta_a}{1+c} \text{ and } A_{t+1}=a \text{ and } N_t(a)>u\Big\} &\leqslant T\exp\Big(-\frac{2(1+c)^2\log(T)}{2\Delta_a^2}\big(\frac{\Delta_a}{1+c}\big)^2\Big)\\ &\leqslant 1\,.\end{aligned}$$

Step 3. The second sum in (1.6), can be rewritten using the notations of Section 1.2. By introducing the stopping times $\tau_{a,k}$ and the random variables $\widetilde{X}_{a,k} = Y_{\tau_{a,k}}$, then on the event $\big\{N_t(a) = k\big\}$, we have the rewriting

$$\widehat{\mu}_{a,N_t(a)} = \widetilde{\mu}_{a,k} \qquad \text{where} \qquad \widetilde{\mu}_{a,k} = \frac{1}{k}\sum_{j=1}^{k} \widetilde{X}_{a,j}\,.$$

Using these notations, we resort to Hoeffding's inequality, whose application is legitimate since $\varepsilon < \frac{c\Delta_a}{1+c}$; the second sum in (1.6) is bounded by

$$\begin{aligned}\sum_{t=1}^{T-1} \mathbb{P}\Big\{\widehat{\mu}_{a,N_t(a)} - \mu_a \geqslant \frac{c\Delta_a}{1+c} - \varepsilon \;\text{ and }\; A_{t+1} = a\Big\} &= \sum_{k\geqslant 1} \mathbb{P}\Big\{\widetilde{\mu}_{a,k} - \mu_a \geqslant \frac{c\Delta_a}{1+c} - \varepsilon\Big\} \\ &\leqslant \sum_{k\geqslant 1} \exp\big(-2k(\frac{c\Delta_a}{1+c} - \varepsilon)^2\big)\,.\end{aligned}$$

Step 4. The first term in (1.6)**.** We define $Q = \ulcorner 2/\varepsilon \urcorner$ and introduce for all $1 \leqslant k \leqslant T$ and $1 \leqslant q \leqslant Q$ the three following events

$$E_{t,k} = \Big\{\mu^\star - \varepsilon > \max_{a\in\mathcal{A}} \widehat{\mu}_{a,N_t(a)} \text{ and } N_t(a^\star) = k\Big\}\,,\quad E_{k,q} = \Big\{\mu^\star - \widetilde{\mu}_{a^\star,k} \in \Big(\tfrac{q\varepsilon}{2}, \tfrac{(q+1)\varepsilon}{2}\Big]\Big\}\,,$$

and finally $E_{t,k,q} = E_{t,k} \cap E_{k,q}$; the first term in (1.6) is thus bounded by

$$\begin{aligned}\sum_{t=1}^{T-1} \mathbb{P}\Big\{\mu^\star - \varepsilon > \max_{a\in\mathcal{A}} \widehat{\mu}_{a,N_t(a)} \text{ and } A_{t+1} = a\Big\} &= \sum_{q=1}^{Q}\sum_{k=1}^{T-1} \mathbb{E}\Big\{\sum_{t=1}^{T-1} \mathbb{I}_{\{E_{t,k,q} \text{ and } A_{t+1}=a\}}\Big\} \\ &= \sum_{q=1}^{Q}\sum_{k=1}^{T-1}\sum_{m=0}^{T-1} \mathbb{P}\Big\{\sum_{t=1}^{T-1} \mathbb{I}_{\{E_{t,k,q} \text{ and } A_{t+1}=a\}} \geqslant m\Big\}\,,\end{aligned}$$

where we used the fact the term under the expectation sign is a positive random variable in the last line. We focus on the probability term; using the stopping times introduce in step 3, we get

$$\begin{aligned}\mathbb{P}\Big\{\sum_{t=1}^{T-1} \mathbb{I}_{\{E_{t,k,q} \text{ and } A_{t+1}=a\}} \geqslant m\Big\} &= \sum_{(s_{k,l})_{l\leqslant m}} \mathbb{P}\Big\{\bigcap_{l=1}^{m}\big\{E_{s_{k,l},k,q} \text{ and } \tau_{a^\star,k} = s_{k,l} \text{ and } A_{s_{k,l}+1} = a\big\}\Big\} \\ &= \mathbb{P}(E_{k,q}) \sum_{(s_{k,l})_{l\leqslant m}} \mathbb{P}\Big\{\bigcap_{l=1}^{m}\big\{E_{s_{k,l},k} \text{ and } \tau_{a^\star,k} = s_{k,l} \text{ and } A_{s_{k,l}+1} = a\big\} \,|\, E_{k,q}\Big\}\,.\end{aligned}$$

We now apply the chain rule $\mathbb{P}\Big\{\bigcap_{l=1}^{m} B_l\Big\} = \prod_{i=1}^{m} \mathbb{P}\Big\{B_i \,|\, \bigcap_{l=1}^{i-1} B_l\Big\}$ first to the events

$$B_l \overset{\text{def}}{=} \big\{E_{s_{k,l},k} \text{ and } \tau_{a^\star,k} = s_{k,l} \text{ and } A_{s_{k,l}+1} = a\big\}\,,$$

and then to the elements of each B_l; we get

$$\mathbb{P}\Big\{B_i \,|\, \bigcap_{l=1}^{i-1} B_l \cap E_{k,q}\Big\} \;=\; \mathbb{P}\Big\{A_{s_{k,i}+1}=a \,|\, \bigcap_{l=1}^{i-1} B_l \text{ and } E_{s_{k,i},k} \text{ and } \tau_{a^\star,k}=s_{k,i}\Big\}\times \mathbb{P}\Big\{\tau_{a^\star,k}=s_{k,i} \text{ and } E_{s_{k,i},k} \;|\; \bigcap_{l=1}^{i-1} B_l \cap E_{k,q}\Big\}.$$

Finally under the event $\Big\{\bigcap_{l=1}^{i-1} B_l \text{ and } E_{s_{k,i},k} \text{ and } \tau_{a^\star,k}=s_{k,i}\Big\}$, we have the property that $\widehat{\Delta}_{a^\star,s_{k,i}} \leqslant \mu^\star - \varepsilon - \widetilde{\mu}_{a^\star,k} \leqslant (q+1)\varepsilon/2 - \varepsilon$. Thus for all $i \leqslant m$ we get, by definition of the algorithm,

$$p_{s_{k,i}}(a^\star) \;\geqslant\; \frac{1}{|\mathcal{A}|}\exp(-2N_{s_{k,i}}(a^\star)\widehat{\Delta}^2_{a,s_{k,i}}) \;\geqslant\; \frac{1}{|\mathcal{A}|}\exp(-2k((q-1)\varepsilon/2)^2).$$

So far we have proved that

$$\sum_{m=1}^{T-1}\mathbb{P}\Big\{\sum_{t=1}^{T-1}\mathbb{I}_{\{E_{t,k,q} \text{ and } A_{t+1}=a\}} \geqslant m\Big\} \;\leqslant\; \sum_{m=1}^{T-1}\mathbb{P}(E_{k,q})(1-\frac{1}{|\mathcal{A}|}\exp(-2k((q-1)\varepsilon/2)^2))^m \times \sum_{(s_{k,l})_{l\leqslant m}}\prod_{i=1}^{m}\mathbb{P}\Big\{\tau_{a^\star,k}=s_{k,i} \text{ and } E_{s_{k,i},k} \,|\, \bigcap_{l=1}^{i-1} B_l \text{ and } E_{k,q}\Big\},$$

where the last sum in this expression is upper-bounded by 1, and where the term $\mathbb{P}(E_{k,q})$ is bounded by $\exp(-2k(q\varepsilon/2)^2)$ by application of Hoeffding's inequality.

Thus, all in all, we get the following inequalities

$$\begin{aligned}\sum_{t=1}^{T-1}\mathbb{P}\Big\{\mu^\star - \varepsilon > \max_{a\in\mathcal{A}}\widehat{\mu}_{a,N_t(a)} \text{ and } A_{t+1}=a\Big\} &\leqslant |\mathcal{A}|\sum_{q=1}^{Q}\sum_{k=1}^{T-1}\exp(-2k(q\varepsilon/2)^2)\exp(2k((q-1)\varepsilon/2)^2)\\ &\leqslant \frac{|\mathcal{A}|}{(1-\exp(-\varepsilon^2/2))(1-\exp(-\varepsilon^2))}.\end{aligned}$$

Step 5. We conclude by combining the three sums of equation (1.6) together.

$$\mathbb{E}\big[N_T(a)\big] \leqslant \frac{(1+c)^2\log(T)}{2\Delta_a^2}+\frac{1}{|\mathcal{A}|}+1+\frac{1}{1-\exp\big(-2(\frac{c\Delta_a}{1+c}-\varepsilon)^2\big)}+\frac{|\mathcal{A}|}{(1-\exp(-\varepsilon^2/2))(1-\exp(-\varepsilon^2))}.$$

Then, since this is valid for all $\varepsilon < \frac{c\Delta_a}{1+c}$, by setting $\varepsilon \overset{\text{def}}{=} \frac{c\Delta_a}{2(1+c)}$ and by using the fact that for all $u \geqslant 0$ one has $e^{-u} \leqslant 1-u+u^2/2$, and $1-e^{-u} \geqslant u/2$ for $u \in [0,1]$, we deduce that

$$\begin{aligned}\mathbb{E}\big[N_T(a)\big] &\leqslant \frac{(1+c)^2\log(T)}{2\Delta_a^2}+\frac{1}{|\mathcal{A}|}+1+\frac{1}{(\frac{c\Delta_a}{1+c}-\varepsilon)^2}+\frac{8|\mathcal{A}|}{\varepsilon^4}\\ &\leqslant \frac{(1+c)^2\log(T)}{2\Delta_a^2}+\frac{1}{|\mathcal{A}|}+1+\frac{4(1+c)^2}{c^2\Delta_a^2}\big[1+\frac{32|\mathcal{A}|(1+c)^2}{c^2\Delta_a^2}\big].\end{aligned}$$

□

4 Limitations of the bandit setting

In this section, we wonder to which extent the bandit approach may be used to handle quite difficult problems, when the player is not facing a set of arms with fix laws, but evolving laws.

A puzzling example. Actually, let us consider the following adversarial bandit problem with two actions $\mathcal{A} = \{0, 1\}$ suggested by Peter Auer, and defined by a simple automaton with two states, deterministic transitions and rewards. More precisely, if action a_t is played at time t, then we define $r_t(a_t) = r(a_t, a_{t-1})$, where r is defined by $r(0, 0) = 1/3$, $r(1, 0) = 0$, $r(0, 1) = 2/3$ and finally $r(1, 1) = 1/2$. It is summarized in Figure 1.9.

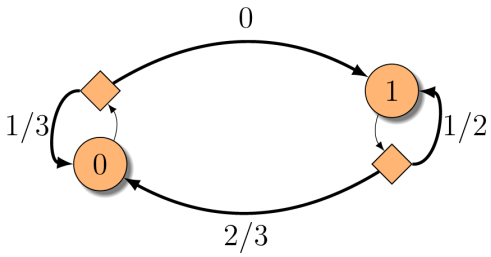


Figure 1.9: A puzzling example.

For such an opponent, the optimal constant strategy in the sense of minimizing (external) regret in hindsight is to play 0, since $1/3 > 0$, and $2/3 > 1/2$. But on the other hand, the constant strategy that always plays 1 achieves higher cumulative reward. Thus minimizing the regret in hindsight differs from maximizing the cumulative reward. What does that mean ?

First, note that even worse, we can build similar examples for which the constant strategy that maximizes the cumulative reward actually maximizes the regret, and the constant strategy that minimizes the regret also minimizes the cumulative reward. Thus, everything seems to go wrong.

Failure of the bandit approach. Let us try to understand which appropriate notion of regret would enable to maximize the cumulative reward. Actually, it appears that learning in hindsight corresponds to learning with a horizon (or look-ahead) 1. Indeed, at time t, we consider the history $h_{<t}$ and try to figure out which action may lead to maximal cumulative reward from this point, but we do not look the consequences if we had played a different action at time $t - 1$. In the previous example, this is what prevents us to play constantly action 1, since we need to consider a horizon of at least 3 to figure out that constantly playing 1 is better than 0. Thus, in order to maximize the cumulative reward in such a problem, we need to consider other definition than the regret in hindsight, which means that the notion of bandit itself is not suitable for such a goal. What we can do is to see this problem from a

reinforcement learning perspective, for which the time horizon is greater than one (depending on the discount factor λ). Such an approach would solve our problem. This shows a limitation of the bandit approach, and at the same time that regret in hindsight may not be always a good thing. But this is not the end of the story.

Failure of the reinforcement learning approach. Even worse, one could think that such a problem will not appear with a reinforcement learning point of view. Actually, it is easy to build a similar problem for which we need a look-ahead horizon n to learn that action 1 gives higher cumulative reward than action 0. More precisely, we can choose the rewards so that after we have played one 0, playing a sequence of ones of length s for any $s < n$ always gives lower cumulative reward than playing anything else for the same amount of time. This means that, since we need to choose action 0 at some point, i.e. explore, to be consistent, we will *experience* the seemingly worst strategy for a possible long term, thus an algorithm has to be fairly confident about the required horizon: you have to accept playing for $n-1$ steps an action that gives a low immediate reward before seeing the benefit of such a strategy. Without prior knowledge of the required look-ahead, we need to consider an infinite look-ahead; note that considering a discounted reinforcement learning does not help much and only undiscounted reinforcement learning can handle such an issue in general, which is a hard setting.

This remark raises the philosophical question to know whether it is always good to try to maximize the cumulative reward at the price of *experiencing* very bad regret (i.e. in hindsight). Indeed finding the strategy that provides the best cumulative reward whatsoever is difficult and may not be possible in the general case since there are classes of opponent that are just not learnable (see Ryabko and Hutter (2008)), whereas on the other hand, all classes are learnable from a bandit perspective, i.e. when we consider the regret in hindsight, making use of algorithms such as Exp3. A better approach would be to build an algorithm that is adaptive to the smallest look-ahead that makes a given problem learnable, but such a difficult question is not currently addressed, for there are plenty of other questions to answer before, like for instance in the case of discounted reinforcement learning setting.

Chapter 2

A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences.

In this chapter, we analyze the stochastic multi-armed bandit setting and more precisely the gap that appears between, on the one hand, the distribution-dependent asymptotic bounds that were derived in Lai and Robbins (1985) and later in Burnetas and Katehakis (1996), which makes appear the Kullback-Leibler divergence between distributions of arms, and, on the other hand, the distribution-dependent non-asymptotic bounds that were derived for UCB-like algorithms in Auer et al. (2002) and later in Audibert et al. (2009),Audibert and Bubeck (2010) or Auer and Ortner (2010), which that makes appear only first or second moments of the distributions of arms.

The former bounds match the asymptotic lower bound, but are only asymptotic and until recently only hold for specific classes of distributions like finite-dimensional parametric distributions - the asymptotic result has been generalized in Honda and Takemura (2010a) for arbitrary classes of distributions; while the later bounds using the UCB algorithms are non-asymptotic and hold for arbitrary distributions (with support included in $[0, 1]$). Moreover, as mentioned in experimental studies by Filippi (2010),Honda and Takemura (2010a) or Garivier and Cappé (2011), the Kullback-Leibler-based algorithms experimentally achieve significantly better performance than the UCB-like algorithms. Unfortunately these bounds do not match the asymptotic lower bounds.

We partially fill this gap by considering a Kullback-Leibler-based algorithm for the stochastic multi-armed bandit problem in the case of distributions with finite support, whose asymptotic regret matches the lower bound of Burnetas and Katehakis (1996), and by providing a finite-time analysis of this algorithm.

This work is a joint work with *Gilles Stoltz*[1], with whom it is very pleasant to work, and has been accepted for publication in the *24th annual Conference On Learning Theory (COLT 2011)*, see Maillard et al. (2011). I also wish to thank *Peter Auer* for insightful discussions while he visited the SequeL Team, and *Daniil Ryabko* for regular discussions about this work.

[1]École normale supérieure, Paris & HEC, Paris

Contents

1 Introduction

The *stochastic* multi-armed bandit problem, introduced by Robbins (1952), formalizes the problem of decision-making under uncertainty, and illustrates the fundamental tradeoff that appears between *exploration*, i.e., making decisions in order to improve the knowledge of the environment, and *exploitation*, i.e., maximizing the payoff.

Setting. In this chapter, we consider a multi-armed bandit problem with *finitely* many arms indexed by $\mathcal{A}$, for which each arm $a \in \mathcal{A}$ is associated with an unknown and fixed probability distribution ν_a over $[0,1]$. The game is *sequential* and goes as follows: at each round $t \geqslant 1$, the player first picks an arm $A_t \in \mathcal{A}$ and then receives a stochastic payoff Y_t drawn at random according to ν_{A_t}. He only gets to see the payoff Y_t.

For each arm $a \in \mathcal{A}$, we denote by μ_a the expectation of its associated distribution ν_a and we let $a^\star$ be any optimal arm, i.e., $a^\star \in \operatorname*{argmax}_{a \in \mathcal{A}} \mu_a$.
We write $\mu^\star$ as a short-hand notation for the largest expectation $\mu_{a^\star}$ and denote the gap of the expected payoff μ_a of an arm $a \in \mathcal{A}$ to $\mu^\star$ as $\Delta_a = \mu^\star - \mu_a$. In addition, the number of times each arm $a \in \mathcal{A}$ is pulled between the rounds 1 and T is referred to as $N_T(a)$,

$$N_T(a) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}} \, .$$

The quality of a strategy will be evaluated through the standard notion of *expected regret*, which we recall now. The expected regret (or simply regret) at round $T \geqslant 1$ is defined as

$$R_T \stackrel{\text{def}}{=} \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} Y_t\right] = \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} \mu_{A_t}\right] = \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}\big[N_T(a)\big] \,, \tag{2.1}$$

where we used the tower rule for the first equality. Note that the expectation is with respect to the random draws of the Y_t according to the ν_{A_t} and also to the possible auxiliary randomizations that the decision-making strategy is resorting to.

The regret measures the cumulative loss resulting from pulling sub-optimal arms, and thus quantifies the amount of exploration required by an algorithm in order to find a best arm, since, as (2.1) indicates, the regret scales with the expected number of pulls of sub-optimal arms. Since the formulation of the problem by Robbins (1952) the regret has been a popular criterion for assessing the quality of a strategy.

Known lower bounds. Lai and Robbins (1985) showed that for some (one-dimensional) parametric classes of distributions, any consistent strategy (i.e., any strategy not pulling sub-optimal arms more than in a polynomial number of rounds) will despite all asymptotically pull in expectation any sub-optimal arm a at least

$$\mathbb{E}\big[N_T(a)\big] \geqslant \left(\frac{1}{\mathcal{K}(\nu_a, \nu^\star)} + o(1)\right) \log(T)$$

times, where $\mathcal{K}(\nu_a, \nu^\star)$ is the Kullback-Leibler (KL) divergence between ν_a and $\nu^\star$; it measures how close distributions ν_a and $\nu^\star$ are from a theoretical information perspective.

Later, Burnetas and Katehakis (1996) extended this result to some classes of multi-dimensional parametric distributions and proved the following generic lower bound: for a given family $\mathcal{P}$ of possible distributions over the arms,

$$\mathbb{E}\big[N_T(a)\big] \geqslant \left(\frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + o(1)\right) \log(T) \,, \text{ where } \quad \mathcal{K}_{\inf}(\nu_a, \mu^\star) \stackrel{\text{def}}{=} \inf_{\nu \in \mathcal{P} \colon E(\nu) > \mu^*} \mathcal{K}(\nu_a, \nu) \,,$$

with the notation $E(\nu)$ for the expectation of a distribution ν. The intuition behind this improvement is to be related to the goal that we want to achieve in bandit problems; it is not detecting whether a distribution is optimal or not (for this goal, the relevant quantity would be $\mathcal{K}(\nu_a, \nu^\star)$), but rather achieving the optimal rate of reward $\mu^\star$ (i.e., one needs to measure how close ν_a is to any distribution $\nu \in \mathcal{P}$ whose expectation is at least $\mu^\star$).

Known upper bounds. Lai and Robbins (1985) provided an algorithm based on the KL divergence, which has been extended by Burnetas and Katehakis (1996) to an algorithm based on $\mathcal{K}_{\inf}$; it is asymptotically optimal since the number of pulls of any sub-optimal arm a satisfies

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

This result holds for finite-dimensional parametric distributions under some assumptions, e.g., the distributions having a finite and known support or belonging to a set of Gaussian distributions with known variance. Recently Honda and Takemura (2010a) extended this asymptotic result to the case of distributions $\mathcal{P}$ with support in $[0,1]$ and such that $\mu^* < 1$; the key ingredient in this case is that $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$ is equal to $\mathcal{K}_{\min}(\nu_a, \mu^\star) \stackrel{\text{def}}{=} \inf_{\nu \in \mathcal{P}: E(\nu) \geqslant \mu^*} \mathcal{K}(\nu_a, \nu)$.

Motivation. All the results mentioned above provide asymptotic bounds only. However, any algorithm is only used for a finite number of rounds and it is thus essential to provide a finite-time analysis of its performance. Auer et al. (2002) initiated this work by providing an algorithm (UCB1) based on a Chernoff-Hoeffding bound; it pulls any sub-optimal arm, till any time T, at most $(8/\Delta_a^2)\log T + 1 + \pi^2/3$ times, in expectation. Although this yields a logarithmic regret, the multiplicative constant depends on the gap $\Delta_a^2 = (\mu^\star - \mu_a)^2$ but not on $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$, which can be seen to be larger than $\Delta_a^2/2$ by Pinsker's inequality; that is, this non-asymptotic bound does not have the right dependence in the distributions. (How much is gained of course depends on the specific families of distributions at hand.) Audibert et al. (2009) provided an algorithm (UCB-V) that takes into account the empirical variance of the arms and exhibited a strategy such that $\mathbb{E}\big[N_T(a)\big] \leqslant 10(\sigma_a^2/\Delta_a^2 + 2/\Delta_a)\log T$ for any time T (where σ_a^2 is the variance of arm a); it improves over UCB1 in case of arms with small variance. Other variants include the MOSS algorithm by Audibert and Bubeck (2010) and Improved UCB by Auer and Ortner (2010).

However, all these algorithms only rely on one moment (for UCB1) or two moments (for UCB-V) of the empirical distributions of the obtained rewards; they do not fully exploit the empirical distributions. As a consequence, the resulting bounds are expressed in terms of the means μ_a and variances σ_a^2 of the sub-optimal arms and not in terms of the quantity $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$ appearing in the lower bounds. The numerical experiments reported in Filippi (2010) confirm that these algorithms are less efficient than those based on $\mathcal{K}_{\inf}$.

Our contribution. In this work we analyze a $\mathcal{K}_{\inf}$-based algorithm inspired by the ones studied in Lai and Robbins (1985), Burnetas and Katehakis (1996), Filippi (2010); it indeed takes into account the full empirical distribution of the observed rewards. The analysis is performed (with explicit bounds) in the case of Bernoulli distributions over the arms. Less explicit but finite-time bounds are obtained in the case of finitely supported distributions (whose supports do not need to be known in advance). Finally, we pave the way for handling the case of general finite-dimensional parametric distributions. These results improve on the ones by Burnetas and Katehakis (1996), Honda and Takemura (2010a) since finite-time bounds (implying their asymptotic results) are obtained; and on Auer et al. (2002), Audibert et al. (2009) as the dependency of the main term scales with $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$. The proposed $\mathcal{K}_{\inf}$-based algorithm is also more natural and more appealing than the one presented in Honda and Takemura (2010a).

Recent related works. Since our initial submission of the present work, we got aware of two papers that tackle problems similar to ours. First, a revised version of Honda and Takemura (2010b, personal communication) obtains finite-time regret bounds (with prohibitively

large constants) for a *randomized* (less natural) strategy in the case of distributions with finite supports (also not known in advance). Second, another paper at this conference (Garivier and Cappé, 2011) also deals with the $\mathcal{K}$–strategy which we study in Theorem 2.1; they however do not obtain second-order terms in closed forms as we do and later extend their strategy to exponential families of distributions (while we extend our strategy to the case of distributions with finite supports). On the other hand, they show how the $\mathcal{K}$–strategy can be extended in a straightforward manner to guarantee bounds with respect to the family of all bounded distributions on a known interval; these bounds are suboptimal but improve on the ones of UCB-type algorithms except maybe for UCB-V.

2 Definitions and tools

Let $\mathcal{X}$ be a Polish space; in the next sections, we will consider $\mathcal{X} = \{0,1\}$ or $\mathcal{X} = [0,1]$. We denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$ and equip $\mathcal{P}(\mathcal{X})$ with the distance d induced by the norm $\|\cdot\|$ defined by $\|\nu\| = \sup_{f\in\mathcal{L}} \left|\int_{\mathcal{X}} f \,\mathrm{d}\nu\right|$, where $\mathcal{L}$ is the set of Lipschitz functions over $\mathcal{X}$, taking values in $[-1,1]$ and with Lipschitz constant smaller than 1.

Kullback-Leibler divergence: For two elements ν, $\kappa \in \mathcal{P}(\mathcal{X})$, we write $\nu \ll \kappa$ when ν is absolutely continuous with respect to κ and denote in this case by $\mathrm{d}\nu/\mathrm{d}\kappa$ the density of ν with respect to κ. We recall that the Kullback-Leibler divergence between ν and κ is defined as

$$\mathcal{K}(\nu,\kappa) = \int_{[0,1]} \frac{\mathrm{d}\nu}{\mathrm{d}\kappa} \log \frac{\mathrm{d}\nu}{\mathrm{d}\kappa}\,\mathrm{d}\kappa \quad \text{if } \nu \ll \kappa; \qquad \text{and} \quad \mathcal{K}(\nu,\kappa) = +\infty \quad \text{otherwise.} \tag{2.2}$$

Empirical distribution: We consider a sequence $X_1, X_2, \ldots$ of random variables taking values in $\mathcal{X}$, independent and identically distributed according to a distribution ν. For all integers $t \geqslant 1$, we denote the empirical distribution corresponding to the first t elements of the sequence by

$$\widehat{\nu}_t = \frac{1}{t}\sum_{s=1}^{t} \delta_{X_t}\,.$$

Non-asymptotic Sanov's Lemma: The following lemma follows from a straightforward adaptation of Dinwoodie (1992, Theorem 2.1 and comments on page 372). Details of the proof are provided in the appendix.

Lemma 2.1 *Let $\mathcal{C}$ be an open convex subset of $\mathcal{P}(\mathcal{X})$ such that* $\Lambda(\mathcal{C}) = \inf_{\kappa\in\mathcal{C}} \mathcal{K}(\kappa,\nu) < \infty$. *Then, for all $t \geqslant 1$, one has the property that*

$$\mathbb{P}_\nu\left\{\widehat{\nu}_t \in \overline{\mathcal{C}}\right\} \leqslant e^{-t\Lambda(\overline{\mathcal{C}})}\,,$$

where $\overline{\mathcal{C}}$ is the closure of $\mathcal{C}$.

This lemma should be thought of as a deviation inequality. The empirical distribution converges (in distribution) to ν. Now, if (and only if) ν is not in the closure of $\mathcal{C}$, then $\Lambda(\mathcal{C}) > 0$ and the lemma indicates how unlikely it is that $\widehat{\nu}_t$ is in this set $\overline{\mathcal{C}}$ not containing the limit ν. The probability of interest decreases at a geometric rate, which depends on $\Lambda(\mathcal{C})$.

3 Finite-time analysis for Bernoulli distributions

In this section, we start with the case of Bernoulli distributions. Although this case is a special case of the general results of Section 4, we provide here a complete and self-contained analysis of this case, where, in addition, we are able to provide closed forms for all the terms in the regret bound. Note however that the resulting bound is slightly worse than what could be derived from the general case (for which more sophisticated tools are used). This result is mainly provided as a warm-up.

3.1 Reminder of some useful results for Bernoulli distributions

We denote by $\mathcal{B}$ the subset of $\mathcal{P}\big([0,1]\big)$ formed by the Bernoulli distributions; it corresponds to $\mathcal{B} = \mathcal{P}\big(\{0,1\}\big)$. A generic element of $\mathcal{B}$ will be denoted by $\beta(p)$, where $p \in [0,1]$ is the probability mass put on 1. We consider a sequence $X_1, X_2, \ldots$ of independent and identically distributed random variables, with common distribution $\beta(p)$; for the sake of clarity we will index, in this subsection only, all probabilities and expectations with p.

For all integers $t \geqslant 1$, we denote by $\quad \widehat{p}_t = \dfrac{1}{t}\displaystyle\sum_{s=1}^{t} X_t \quad$ the empirical average of the first t elements of the sequence.

The lemma below follows from an adaptation of Garivier and Leonardi (2010, Proposition 2). The details of the adaptation (and simplification) can be found in the appendix.

Lemma 2.2 *For all $p \in [0,1]$, all $\varepsilon > 1$, and all $t \geqslant 1$,*

$$\mathbb{P}_p\left(\bigcup_{s=1}^{t}\left\{ s\ \mathcal{K}\Big(\beta\big(\widehat{p}_s\big),\ \beta(p)\Big) \geqslant \varepsilon \right\}\right) \leqslant 2e\,\big\lceil \varepsilon \log t \big\rceil\, e^{-\varepsilon}\,.$$

In particular, for all random variables N_t taking values in $\{1,\ldots,t\}$,

$$\mathbb{P}_p\left\{ N_t\ \mathcal{K}\Big(\beta\big(\widehat{p}_{N_t}\big),\ \beta(p)\Big) \geqslant \varepsilon \right\} \leqslant 2e\,\big\lceil \varepsilon \log t \big\rceil\, e^{-\varepsilon}\,.$$

Another immediate fact about Bernoulli distributions is that for all $p \in (0,1)$, the mappings $\mathcal{K}_{\cdot,p} : q \in (0,1) \mapsto \mathcal{K}\big(\beta(p),\beta(q)\big)$ and $\mathcal{K}_{p,\cdot} : q \in [0,1] \mapsto \mathcal{K}\big(\beta(q),\beta(p)\big)$ are continuous and take finite values. In particular, we have, for instance, that for all $\varepsilon > 0$ and $p \in (0,1)$, the set

$$\Big\{ q \in [0,1] :\quad \mathcal{K}\big(\beta(p),\beta(q)\big) \leqslant \varepsilon \Big\}$$

is a closed interval containing p. This property still holds when $p \in \{0,1\}$, as in this case, the interval is reduced to $\{p\}$.

3.2 Strategy and analysis

We consider the so-called $\mathcal{K}$*–strategy* of Figure 2.1, which was already considered in the literature, see Burnetas and Katehakis (1996), Filippi (2010). The numerical computation of the quantities $B^+_{a,t}$ is straightforward (by convexity of $\mathcal{K}$ in its second argument, by using iterative methods) and is detailed therein.

Parameters: A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

Initialization: Pull each arm of $\mathcal{A}$ once

For rounds $t+1$, where $t \geqslant |\mathcal{A}|$,

- compute for each arm $a \in \mathcal{A}$ the quantity

$$B^+_{a,t} = \max\left\{ q \in [0,1] : \quad N_t(a)\, \mathcal{K}\Big(\beta\big(\widehat{\mu}_{a,N_t(a)}\big),\, \beta(q)\Big) \leqslant f(t) \right\},$$

  where $$\widehat{\mu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} Y_s \,;$$

- in case of a tie, pick an arm with largest value of $\widehat{\mu}_{a,N_t(a)}$;

- pull any arm $A_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} B^+_{a,t}$.

Figure 2.1: The $\mathcal{K}$–strategy.

Before proceeding, we denote by $\sigma_a^2 = \mu_a(1-\mu_a)$ the variance of each arm $a \in \mathcal{A}$ (and take the short-hand notation $\sigma^{\star,2}$ for the variance of an optimal arm).

Theorem 2.1 (*Regret bound for the $\mathcal{K}$-strategy*) *When $\mu^\star \in (0,1)$, for all non-decreasing functions $f : \mathbb{N} \to \mathbb{R}_+$ such that $f(1) \geqslant 1$, the expected regret R_T of the strategy of Figure 2.1 is upper bounded by the infimum, as the $(c_a)_{a\in\mathcal{A}}$ describe $(0,+\infty)$, of the quantities*

$$\sum_{a \in \mathcal{A}} \Delta_a \left(\frac{(1+c_a)\, f(T)}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} + 4e \sum_{t=|\mathcal{A}|}^{T-1} \big\lceil f(t) \log t \big\rceil\, e^{-f(t)} + \frac{(1+c_a)^2}{8\, c_a^2 \Delta_a^2\, \min\big\{\sigma_a^4,\, \sigma^{\star,4}\big\}} \mathbb{I}_{\{\mu_a \in (0,1)\}} + 3 \right).$$

For $\mu^\star = 0$, its regret is null. For $\mu^\star = 1$, it satisfies $R_T \leqslant 2\big(|\mathcal{A}| - 1\big)$.

A possible choice for the function f is $f(t) = \log\big((et)\log^3(et)\big)$, which is non decreasing, satisfies $f(1) \geqslant 1$, and is such that the second term in the sum above is bounded (by a

basic result about so-called Bertrand's series). Now, as the constants c_a in the bound are parameters of the analysis (and not of the strategy), they can be optimized. For instance, with the choice of $f(t)$ mentioned above, taking each c_a proportional to $(\log T)^{-1/3}$ (up to a multiplicative constant that depends on the distributions ν_a) entails the regret bound

$$\sum_{a\in\mathcal{A}} \Delta_a \frac{\log T}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} + \varepsilon_T\,,$$

where it is easy to give an explicit and closed-form expression of ε_T; in this conference version, we only indicate that ε_T is of order of $(\log T)^{2/3}$ but we do not know whether the order of magnitude of this second-order term is optimal.

Proof: We first deal with the case where $\mu^\star \notin \{0,1\}$ and introduce an additional notation. In view of the remark at the end of Section 3.1, for all arms a and rounds t, we let $B^-_{a,t}$ be the element in $[0,1]$ such that

$$\Big\{q\in[0,1]:\quad N_t(a)\;\mathcal{K}\Big(\beta\big(\widehat{\mu}_{a,N_t(a)}\big),\,\beta(q)\Big)\leqslant f(t)\Big\} = \big[B^-_{a,t},\; B^+_{a,t}\big]\,. \tag{2.3}$$

As (2.1) indicates, it suffices to bound $N_T(a)$ for all suboptimal arms a, i.e., for all arms such that $\mu_a<\mu^\star$. We will assume in addition that $\mu_a>0$ (and we also have $\mu_a\leqslant\mu^\star<1$); the case where $\mu_a=0$ will be handled separately.

Step 1: A decomposition of the events of interest. For $t\geqslant|\mathcal{A}|$, when $A_{t+1}=a$, we have in particular, by definition of the strategy, that $B^+_{a,t}\geqslant B^+_{a^\star,t}$. On the event

$$\big\{A_{t+1}=a\big\}\cap\Big\{\mu^\star\in\big[B^-_{a^\star,t},\; B^+_{a^\star,t}\big]\Big\}\cap\Big\{\mu_a\in\big[B^-_{a,t},\; B^+_{a,t}\big]\Big\}\,,$$

we therefore have, on the one hand, $\mu^\star\leqslant B^+_{a^\star,t}\leqslant B^+_{a,t}$ and on the other hand, $B^-_{a,t}\leqslant\mu_a\leqslant\mu^\star$, that is, the considered event is included in $\Big\{\mu^\star\in\big[B^-_{a,t},\; B^+_{a,t}\big]\Big\}$. We thus proved that

$$\big\{A_{t+1}=a\big\}\subseteq\Big\{\mu^\star\notin\big[B^-_{a^\star,t},\; B^+_{a^\star,t}\big]\Big\}\cup\Big\{\mu_a\notin\big[B^-_{a,t},\; B^+_{a,t}\big]\Big\}\cup\Big\{\mu^\star\in\big[B^-_{a,t},\; B^+_{a,t}\big]\Big\}\,.$$

Going back to the definition (2.3), we get in particular the inclusion

$$\begin{aligned}\big\{A_{t+1}=a\big\}\subseteq\quad&\Big\{N_t(a^\star)\;\mathcal{K}\Big(\beta\big(\widehat{\mu}_{a^\star,N_t(a^\star)}\big),\,\beta(\mu^\star)\Big)>f(t)\Big\}\\ &\cup\Big\{N_t(a)\;\mathcal{K}\Big(\beta\big(\widehat{\mu}_{a,N_t(a)}\big),\,\beta(\mu_a)\Big)>f(t)\Big\}\\ &\cup\left(\Big\{N_t(a)\;\mathcal{K}\Big(\beta\big(\widehat{\mu}_{a,N_t(a)}\big),\,\beta(\mu^\star)\Big)\leqslant f(t)\Big\}\cap\big\{A_{t+1}=a\big\}\right).\end{aligned}$$

Step 2: Bounding the probabilities of two elements of the decomposition. We consider the filtration $(\mathcal{F}_t)$, where for all $t\geqslant1$, the σ–algebra $\mathcal{F}_t$ is generated by $A_1,Y_1,\ldots,$

A_t, Y_t. In particular, A_{t+1} and thus all $N_{t+1}(a)$ are $\mathcal{F}_t$–measurable. We denote by $\tau_{a,1}$ the deterministic round at which a was pulled for the first time and by $\tau_{a,2}$, $\tau_{a,3}$, ... the rounds $t \geqslant |\mathcal{A}| + 1$ at which a was then played; since for all $k \geqslant 2$,

$$\tau_{a,k} = \min\big\{t \geqslant |\mathcal{A}| + 1 : \quad N_t(a) = k\big\}\,,$$

we see that $\big\{\tau_{a,k} = t\big\}$ is $\mathcal{F}_{t-1}$–measurable. Therefore, for each $k \geqslant 1$, the random variable $\tau_{a,k}$ is a (predictable) stopping time. Hence, by a well-known fact in probability theory (see, e.g., Chow and Teicher 1988, Section 5.3), the random variables $\widetilde{X}_{a,k} = Y_{\tau_{a,k}}$, where $k = 1, 2, \ldots$ are independent and identically distributed according to ν_a. Since on $\big\{N_t(a) = k\big\}$, we have the rewriting

$$\widehat{\mu}_{a,N_t(a)} = \widetilde{\mu}_{a,k} \qquad \text{where} \qquad \widetilde{\mu}_{a,k} = \frac{1}{k}\sum_{j=1}^{k} \widetilde{X}_{a,j}\,,$$

and since for $t \geqslant |\mathcal{A}| + 1$, one has $N_t(a) \geqslant 1$ with probability 1, we can apply the second statement in Lemma 2.2 and get, for all $t \geqslant |\mathcal{A}| + 1$,

$$\mathbb{P}\bigg\{N_t(a)\ \mathcal{K}\Big(\beta\big(\widehat{\mu}_{a,N_t(a)}\big),\, \beta(\mu_a)\Big) > f(t)\bigg\} \leqslant 2e\,\big\lceil f(t)\log t\big\rceil\, e^{-f(t)}\,.$$

A similar argument shows that for all $t \geqslant |\mathcal{A}| + 1$,

$$\mathbb{P}\bigg\{N_t(a^\star)\ \mathcal{K}\Big(\beta\big(\widehat{\mu}_{a^\star,N_t(a^\star)}\big),\, \beta(\mu^\star)\Big) > f(t)\bigg\} \leqslant 2e\,\big\lceil f(t)\log t\big\rceil\, e^{-f(t)}\,.$$

Step 3: Rewriting the remaining terms. We therefore proved that

$$\mathbb{E}\big[N_T(a)\big] \leqslant 1 + 4e\sum_{t=|\mathcal{A}|}^{T-1} \big\lceil f(t)\log t\big\rceil\, e^{-f(t)} + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\bigg(\bigg\{N_t(a)\ \mathcal{K}\Big(\beta\big(\widehat{\mu}_{a,N_t(a)}\big),\, \beta(\mu^\star)\Big) \leqslant f(t)\bigg\} \cap \big\{A_{t+1} = a\big\}\bigg)$$

and deal now with the last sum. Since f is non decreasing, it is bounded by

$$\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big(K_t \cap \big\{A_{t+1} = a\big\}\Big) \qquad \text{where} \qquad K_t = \bigg\{N_t(a)\ \mathcal{K}\Big(\beta\big(\widehat{\mu}_{a,N_t(a)}\big),\, \beta(\mu^\star)\Big) \leqslant f(T)\bigg\}\,.$$

Now, $$\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big(K_t \cap \big\{A_{t+1} = a\big\}\Big) = \mathbb{E}\left[\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{I}_{\big\{A_{t+1}=a\big\}}\mathbb{I}_{K_t}\right] = \mathbb{E}\left[\sum_{k\geqslant 2} \mathbb{I}_{\big\{\tau_{a,k}\leqslant T\big\}}\mathbb{I}_{K_{\tau_{a,k}-1}}\right].$$

We note that, since $N_{\tau_{a,k}-1}(a) = k - 1$, we have that

$$K_{\tau_{a,k}-1} = \bigg\{(k-1)\ \mathcal{K}\Big(\beta\big(\widetilde{\mu}_{a,k-1}\big),\, \beta(\mu^\star)\Big) \leqslant f(T)\bigg\}\,.$$

All in all, since $\tau_{a,k} \leqslant T$ implies $k \leqslant T - |\mathcal{A}| + 1$ (since each arm is played at least once during the first $|\mathcal{A}|$ rounds), we have

$$\mathbb{E}\left[\sum_{k\geqslant 2} \mathbb{I}_{\left\{\tau_{a,k}\leqslant T\right\}} \mathbb{I}_{K_{\tau_{a,k}-1}}\right] \leqslant \mathbb{E}\left[\sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{I}_{K_{\tau_{a,k}-1}}\right] = \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\Big\{(k-1)\, \mathcal{K}\Big(\beta\big(\widetilde{\mu}_{a,k-1}\big),\, \beta(\mu^\star)\Big) \leqslant f(T)\Big\}. \tag{2.4}$$

Step 4: Bounding the probabilities of the latter sum via Sanov's lemma. For each $\gamma > 0$, we define the convex open set $\mathcal{C}_\gamma = \Big\{\beta(q) \in \mathcal{B} : \ \mathcal{K}\big(\beta(q),\, \beta(\mu^\star)\big) < \gamma\Big\}$, which is a non-empty set (since $\mu^\star < 1$); by continuity of the mapping $\mathcal{K}_{\cdot,\mu^\star}$ defined after the statement of Lemma 2.2 when $\mu^\star \in (0,1)$, its closure equals $\overline{\mathcal{C}}_\gamma = \Big\{\beta(q) \in \mathcal{B} : \ \mathcal{K}\big(\beta(q),\, \beta(\mu^\star)\big) \leqslant \gamma\Big\}$.

In addition, since $\mu_a \in (0,1)$, we have that $\mathcal{K}\big(\beta(q),\, \beta(\mu_a)\big) < \infty$ for all $q \in [0,1]$. In particular, for all $\gamma > 0$, the condition $\Lambda\big(\mathcal{C}_\gamma\big) < \infty$ of Lemma 2.7 is satisfied. Denoting this value by

$$\theta_a(\gamma) = \inf\Big\{\mathcal{K}\big(\beta(q),\, \beta(\mu_a)\big) : \quad \beta(q) \in \mathcal{B} \ \text{ such that } \ \mathcal{K}\big(\beta(q),\, \beta(\mu^\star)\big) \leqslant \gamma\Big\},$$

we get by the indicated lemma that for all $k \geqslant 1$,

$$\mathbb{P}\Big\{\mathcal{K}\Big(\beta\big(\widetilde{\mu}_{a,k}\big),\, \beta(\mu^\star)\Big) \leqslant \gamma\Big\} = \mathbb{P}\Big\{\beta\big(\widetilde{\mu}_{a,k}\big) \in \overline{\mathcal{C}}_\gamma\Big\} \leqslant e^{-k\,\theta_a(\gamma)}.$$

Now, since (an open neighborhood of) $\beta(\mu_a)$ is not included in $\overline{\mathcal{C}}_\gamma$ as soon as $0 < \gamma < \mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)$, we have that $\theta_a(\gamma) > 0$ for such values of γ. To apply the obtained inequality to the last sum in (2.4), we fix a constant $c_a > 0$ and denote by k_0 the following upper integer part, $k_0 = \left\lceil \dfrac{(1+c_a)\, f(T)}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} \right\rceil$, so that $f(T)/k \leqslant \mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a) < \mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)$ for $k \geqslant k_0$, hence,

$$\begin{aligned}\sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\Big\{(k-1)\, \mathcal{K}\Big(\beta\big(\widetilde{\mu}_{a,k-1}\big),\, \beta(\mu^\star)\Big) \leqslant f(T)\Big\} \ &\leqslant \ \sum_{k=1}^{T} \mathbb{P}\Big\{\mathcal{K}\Big(\beta\big(\widetilde{\mu}_{a,k}\big),\, \beta(\mu^\star)\Big) \leqslant \frac{f(T)}{k}\Big\} \\ &\leqslant \ k_0 - 1 + \sum_{k=k_0}^{T} \exp\Big(-k\,\theta_a\big(f(T)/k\big)\Big).\end{aligned}$$

Since θ_a is a non-increasing function,

$$\begin{aligned}\sum_{k=k_0}^{T} \exp\Big(-k\,\theta_a\big(f(T)/k\big)\Big) \ &\leqslant \ \sum_{k=k_0}^{T} \exp\Big(-k\,\theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)\Big) \\ &\leqslant \ \Gamma_a(c_a)\, \exp\Big(-k_0\,\theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)\Big) \leqslant \Gamma_a(c_a),\end{aligned}$$

where $\Gamma_a(c_a) = \Big[1 - \exp\Big(-\theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)\Big)\Big]^{-1}$.
Putting all pieces together, we thus proved so far that

$$\mathbb{E}\big[N_T(a)\big] \leqslant 1 + \frac{(1+c_a)\, f(T)}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} + 4e \sum_{t=|\mathcal{A}|}^{T-1} \big\lceil f(t) \log t \big\rceil\, e^{-f(t)} + \Gamma_a(c_a)$$

and it only remains to deal with $\Gamma_a(c_a)$.

Step 5: Getting an upper bound in closed form for $\Gamma_a(c_a)$. We will make repeated uses of Pinsker's inequality: for $p, q \in [0,1]$, one has $\mathcal{K}\big(\beta(p), \beta(q)\big) \geqslant 2\,(p-q)^2$.
In what follows, we use the short-hand notation $\Theta_a = \theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)$ and therefore need to upper bound $1/\big(1-e^{-\Theta_a}\big)$. Since for all $u \geqslant 0$, one has $e^{-u} \leqslant 1-u+u^2/2$, we get $\Gamma_a(c_a) \leqslant \dfrac{1}{\Theta_a\big(1-\Theta_a/2\big)} \leqslant \dfrac{2}{\Theta_a}$ for $\Theta_a \leqslant 1$, and $\Gamma_a(c_a) \leqslant \dfrac{1}{1-e^{-1}} \leqslant 2$ for $\Theta_a \geqslant 1$. It thus only remains to lower bound Θ_a in the case when it is smaller than 1.

By the continuity properties of the Kullback-Leibler divergence, the infimum in the definition of θ_a is always achieved; we therefore let $\widetilde{\mu}$ be an element in $[0,1]$ such that

$$\Theta_a = \mathcal{K}\big(\beta(\widetilde{\mu}),\, \beta(\mu_a)\big) \qquad \text{and} \qquad \mathcal{K}\big(\beta(\widetilde{\mu}),\, \beta(\mu^\star)\big) = \frac{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)}{1+c}\,;$$

it is easy to see that we have the ordering $\mu_a < \widetilde{\mu} < \mu^\star$. By Pinsker's inequality, $\Theta_a \geqslant 2\big(\widetilde{\mu} - \mu_a\big)^2$ and we now lower bound the latter quantity. We use the short-hand notation $f(p) = \mathcal{K}\big(\beta(p), \beta(\mu^\star)\big)$ and note that the thus defined mapping f is convex and differentiable on $(0,1)$; its derivative equals

$$f'(p) = \log \frac{1-\mu^\star}{\mu^\star} + \log \frac{p}{1-p}$$

for all $p \in (0,1)$ and is therefore non positive for $p \leqslant \mu^\star$. By the indicated convexity of f, using a subgradient inequality, we get $f\big(\widetilde{\mu}\big) - f(\mu_a) \geqslant f'(\mu_a)\,\big(\widetilde{\mu} - \mu_a\big)$, which entails, since $f'(\mu_a) < 0$,

$$\widetilde{\mu} - \mu_a \geqslant \frac{f\big(\widetilde{\mu}\big) - f(\mu_a)}{f'(\mu_a)} = \frac{c_a}{1+c_a}\; \frac{f(\mu_a)}{-f'(\mu_a)}\,, \tag{2.5}$$

where the equality follows from the fact that by definition of μ, we have $f\big(\widetilde{\mu}\big) = f(\mu_a)/(1+c_a)$. Now, since f' is differentiable as well on $(0,1)$ and takes the value 0 at $\mu^\star$, a Taylor's equality entails that there exists a $\xi \in (\mu_a, \mu^\star)$ such that

$$-f'(\mu_a) = f'(\mu^\star) - f'(\mu_a) = f''(\xi)\,\big(\mu^\star - \mu_a\big) \qquad \text{where} \quad f''(\xi) = \frac{1}{\xi} + \frac{1}{1-\xi} = \frac{1}{\xi(1-\xi)}\,.$$

Therefore, by convexity of $\tau \mapsto \tau(1-\tau)$, we get that

$$\frac{1}{-f'(\mu_a)} \geqslant \frac{\min\big\{\mu_a(1-\mu_a),\, \mu^\star(1-\mu^\star)\big\}}{\mu^\star - \mu_a}\,.$$

Substituting this into (2.5) and using again Pinsker's inequality to lower bound $f(\mu_a)$, we have proved

$$\widetilde{\mu} - \mu_a \geqslant 2\,\frac{c_a}{1+c_a}\,\big(\mu^\star - \mu_a\big)\,\min\big\{\mu_a(1-\mu_a),\,\mu^\star(1-\mu^\star)\big\}\,.$$

Putting all pieces together, we thus proved that

$$\Gamma_a(c_a) \leqslant 2\,\max\left\{\frac{(1+c_a)^2}{8\,c_a^2\big(\mu^\star-\mu_a\big)^2\Big(\min\big\{\mu_a(1-\mu_a),\,\mu^\star(1-\mu^\star)\big\}\Big)^2},\;1\right\}\,;$$

bounding the maximum of the two quantities by their sum concludes the main part of the proof.

Step 6: For $\mu^\star \in \{0,1\}$ and/or $\mu_a = 0$. When $\mu^\star = 1$, then $\widehat{\mu}_{a^\star,N_t(a\star)} = 1$ for all $t \geqslant |\mathcal{A}|+1$, so that $B^+_{a^\star,t} = 1$ for all $t \geqslant |\mathcal{A}|+1$. Thus, the arm a is played after round $t \geqslant |\mathcal{A}|+1$ only if $B^+_{a,t} = 1$ and $\widehat{\mu}_{a,N_t(a)} = 1$ (in view of the tie-breaking rule of the considered strategy). But this means that a is played as long as it gets payoffs equal to 1 and is stopped being played when it receives the payoff 0 for the first time. Hence, in this case, we have that the sum of payoffs equals at least $T - 2\big(|\mathcal{A}|-1\big)$ and the regret $R_T = \mathbb{E}[T\mu^\star - (Y_1+\ldots+Y_t)]$ is therefore bounded by $2\big(|\mathcal{A}|-1\big)$.

When $\mu^\star = 0$, a Dirac mass over 0 is associated with all arms and the regret of all strategies is equal to 0.

We consider now the case $\mu^\star \in (0,1)$ and $\mu_a = 0$, for which the first three steps go through; only in the upper bound of step 4 we used the fact that $\mu_a > 0$. But in this case, we have a deterministic bound on (2.4). Indeed, since $\mathcal{K}\big(\beta(0),\beta(\mu^\star)\big) = -\log\mu^\star$, we have $k\,\mathcal{K}\big(\beta(0),\beta(\mu^\star)\big) \leqslant f(T)$ if and only if

$$k \leqslant \frac{f(T)}{-\log\mu^\star} = \frac{f(T)}{\mathcal{K}\big(\beta(\mu_a),\beta(\mu^\star)\big)}\,,$$

which improves on the general bound exhibited in step 4. □

Remark 1 Note that Step 5 in the proof is specifically designed to provide an upper bound on $\Gamma_a(c_a)$ in the case of Bernoulli distributions. In the general case, getting such an explicit bound seems more involved.

4 A finite-time analysis in the case of distributions with finite support

Before stating and proving our main result, Theorem 2.2, we introduce the quantity $\mathcal{K}_{\inf}$ and list some of its properties.

4.1 Some useful properties of $\mathcal{K}_{\inf}$ and its level sets

We now introduce the key quantity in order to generalize the previous algorithm to handle the case of distributions with finite support. To that end, we introduce $\mathcal{P}_F\big([0,1]\big)$, the subset of $\mathcal{P}\big([0,1]\big)$ that consists of distributions with finite support.

Definition 2.1 *For all distributions $\nu \in \mathcal{P}_F\big([0,1]\big)$ and $\mu \in [0,1)$, we define*

$$\mathcal{K}_{\inf}(\nu,\mu) = \inf\Big\{\mathcal{K}(\nu,\nu') : \quad \nu' \in \mathcal{P}_F\big([0,1]\big) \quad \text{s.t.} \quad E(\nu') > \mu\Big\},$$

where $E(\nu') = \int_{[0,1]} x \,\mathrm{d}\nu'(x)$ denotes the expectation of the distribution ν'.

We now remind some useful properties of $\mathcal{K}_{\inf}$. Honda and Takemura (2010b, Lemma 6) can be reformulated in our context as follows.

Lemma 2.3 *For all $\nu \in \mathcal{P}_F\big([0,1]\big)$, the mapping $\mathcal{K}_{\inf}(\nu,\,\cdot\,)$ is continuous and non decreasing in its argument $\mu \in [0,1)$. Moreover, the mapping $\mathcal{K}_{\inf}(\,\cdot\,,\mu)$ is lower semi-continuous on $\mathcal{P}_F\big([0,1]\big)$ for all $\mu \in [0,1)$.*

The next two lemmas bound the variation of $\mathcal{K}_{\inf}$, respectively in its first and second arguments. (For clarity, we denote the expectations with respect to ν by $\mathbb{E}_\nu$.) Their proofs are both deferred to the appendix. We denote by $\|\,\cdot\,\|_1$ the ℓ^1–norm on $\mathcal{P}\big([0,1]\big)$ and recall that the ℓ^1–norm of $\nu - \nu'$ corresponds to twice the distance in variation between ν and ν'.

Lemma 2.4 *For all $\mu \in (0,1)$ and for all $\nu,\, \nu' \in \mathcal{P}_F\big([0,1]\big)$, the following holds true.*

- *In the case when $\mathbb{E}_\nu\big[(1-\mu)/(1-X)\big] > 1$, then $\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) \leqslant M_{\nu,\mu}\,\|\nu-\nu'\|_1$, for some constant $M_{\nu,\mu} > 0$.*
- *In the case when $\mathbb{E}_\nu\big[(1-\mu)/(1-X)\big] \leqslant 1$, the fact that $\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) \geqslant \alpha\,\mathcal{K}_{\inf}(\nu,\mu)$ for some $\alpha \in (0,1)$ entails that*

$$\|\nu - \nu'\|_1 \geqslant \frac{1-\mu}{(2/\alpha)\,\big((2/\alpha)-1\big)}\,.$$

Lemma 2.5 *We have that for any $\nu \in \mathcal{P}_F\big([0,1]\big)$, provided that $\mu \geqslant \mu - \varepsilon > E(\nu)$, the following inequalities hold true:*

$$\varepsilon/(1-\mu) \geqslant \mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) \geqslant 2\varepsilon^2$$

Moreover, the first inequality is also valid when $E(\nu) \geqslant \mu > \mu - \varepsilon$ or $\mu > E(\nu) \geqslant \mu - \varepsilon$.

Level sets of $\mathcal{K}_{\inf}$: For each $\gamma > 0$ and $\mu \in (0,1)$, we consider the set

$$\begin{aligned}\mathcal{C}_{\mu,\gamma} &= \Big\{\nu' \in \mathcal{P}_F\big([0,1]\big) : \ \mathcal{K}_{\inf}(\nu',\mu) < \gamma\Big\} \\ &= \Big\{\nu' \in \mathcal{P}_F\big([0,1]\big) : \ \exists\, \nu'_\mu \in \mathcal{P}_F\big([0,1]\big) \ \text{ s.t. } \ E\big(\nu'_\mu\big) > \mu \ \text{ and } \ \mathcal{K}\big(\nu',\nu'_\mu\big) < \gamma\Big\}.\end{aligned}$$

We detail a property in the following lemma, whose proof is also deferred to the appendix.

Lemma 2.6 *For all $\gamma > 0$ and $\mu \in (0,1)$, the set $\mathcal{C}_{\mu,\gamma}$ is a non-empty open convex set. Moreover,*

$$\overline{\mathcal{C}}_{\mu,\gamma} \supseteq \Big\{\nu' \in \mathcal{P}_F\big([0,1]\big) : \ \mathcal{K}_{\inf}(\nu',\mu) \leqslant \gamma\Big\}.$$

4.2 The $\mathcal{K}_{\inf}$–strategy and a general performance guarantee

For each arm $a \in \mathcal{A}$ and round t with $N_t(a) > 0$, we denote by $\widehat{\nu}_{a,N_t(a)}$ the empirical distribution of the payoffs obtained till round t when picking arm a, that is,

$$\widehat{\nu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} \delta_{Y_s}\,,$$

where for all $x \in [0,1]$, we denote by δ_x the Dirac mass on x. We define the corresponding empirical averages as

$$\widehat{\mu}_{a^\star,N_t(a^\star)} = E\big(\widehat{\nu}_{a^\star,N_t(a^\star)}\big) = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} Y_s\,.$$

We then consider the $\mathcal{K}_{\inf}$*–strategy* defined in Figure 2.2. Note that the use of maxima in the definitions of the $B^+_{a,t}$ is justified by Lemma 2.3.

As explained in Honda and Takemura (2010b), the computation of the quantities $\mathcal{K}_{\inf}$ can be done efficiently in this case, i.e., when we consider only distributions with finite supports. This is because in the computation of $\mathcal{K}_{\inf}$, it is sufficient to consider only distributions with the same support as the empirical distributions (up to one point). Note that the knowledge of the support of the distributions associated with the arms is not required.

Theorem 2.2 (*Regret bound for the $\mathcal{K}_{\inf}$-strategy*) *Assume that $\nu^\star$ is finitely supported, with expectation $\mu^\star \in (0,1)$ and with support denoted by $\mathcal{S}^\star$. Let $a \in \mathcal{A}$ be a suboptimal arm such that $\mu_a > 0$ and ν_a is finitely supported. Then, for all $c_a > 0$ and all*

$$0 < \varepsilon < \min\left\{\Delta_a,\ \frac{c_a/2}{1+c_a}(1-\mu^\star)\,\mathcal{K}_{\inf}(\nu_a,\mu^\star)\right\},$$

the expected number of times the $\mathcal{K}_{\inf}$–strategy, run with $f(t) = \log t$, pulls arm a satisfies

$$\mathbb{E}\big[N_T(a)\big] \leqslant 1 + \frac{(1+c_a)\,\log T}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)} + \frac{1}{1-e^{-\Theta_a(c_a,\varepsilon)}} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{1-\mu^*+\varepsilon}\right)\sum_{k=1}^{T}(k+1)^{|\mathcal{S}^\star|}\,e^{-k\varepsilon^2} + \frac{1}{(\Delta_a-\varepsilon)^2}\,,$$

Parameters: A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

Initialization: Pull each arm of $\mathcal{A}$ once

For rounds $t+1$, where $t \geqslant |\mathcal{A}|$,

- compute for each arm $a \in \mathcal{A}$ the quantity

$$B_{a,t}^{+} = \max\Big\{q \in [0,1] : \quad N_t(a)\ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,N_t(a)},\, q\big) \leqslant f(t)\Big\}\,,$$

  where $\displaystyle \widehat{\nu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} \delta_{Y_s}\,;$

- in case of a tie, pick an arm with largest value of $\widehat{\mu}_{a,N_t(a)}$;

- pull any arm $A_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} B_{a,t}^{+}$.

Figure 2.2: The strategy $\mathcal{K}_{\inf}$.

where

$$\Theta_a(c_a, \varepsilon) = \theta_a\left(\frac{\log T}{k_0} + \frac{\varepsilon}{1-\mu^\star}\right) \qquad \textit{with} \qquad k_0 = \left\lceil \frac{(1+c_a)\,\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} \right\rceil\,.$$

and for all $\gamma > 0$,

$$\theta_a(\gamma) = \inf\Big\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \ \text{ s.t. } \ \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\Big\}\,.$$

As a corollary, we get (by taking some common value for all c_a) that for all $c > 0$,

$$\overline{R}_T \leqslant \sum_{a \in \mathcal{A}} \Delta_a \frac{(1+c)\,\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + h(c)\,,$$

where $h(c) < \infty$ is a function of c (and of the distributions associated with the arms), which is however independent of T. As a consequence, we recover the asymptotic results of Burnetas and Katehakis (1996), Honda and Takemura (2010a), i.e., the guarantee that

$$\limsup_{T \to \infty} \frac{\overline{R}_T}{\log T} \leqslant \sum_{a \in \mathcal{A}} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

Of course, a sharper optimization can be performed by carefully choosing the constants c_a, that are parameters of the analysis; similarly to the comments after the statement of Theorem 2.1, we would then get a dominant term with a constant factor 1 instead of $1+c$ as above, plus an additional second-order term. Details are left to a journal version of this work.

Proof: By arguments similar to the ones used in the first step of the proof of Theorem 2.1, we have

$$\Big\{A_{t+1}=a\Big\} \subseteq \Big\{\mu^\star-\varepsilon<\widehat{\mu}_{a,N_t(a)}\Big\} \cup \Big\{\mu^\star-\varepsilon>B^+_{a^\star,t}\Big\} \cup \Big\{\mu^\star-\varepsilon\in\big[\widehat{\mu}_{a,N_t(a)},\ B^+_{a,t}\big]\Big\};$$

indeed, on the event $\Big\{A_{t+1}=a\Big\} \cap \Big\{\mu^\star-\varepsilon\geqslant\widehat{\mu}_{a,N_t(a)}\Big\} \cap \Big\{\mu^\star-\varepsilon\leqslant B^+_{a^\star,t}\Big\}$, we have, $\widehat{\mu}_{a,N_t(a)} \leqslant \mu^\star-\varepsilon \leqslant B^+_{a^\star,t} \leqslant B^+_{a,t}$ (where the last inequality is by definition of the strategy). Before proceeding, we note that

$$\Big\{\mu^\star-\varepsilon\in\big[\widehat{\mu}_{a,N_t(a)},\ B^+_{a,t}\big]\Big\} \subseteq \Big\{N_t(a)\ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,N_t(a)},\ \mu^\star-\varepsilon\big)\leqslant f(t)\Big\},$$

since $\mathcal{K}_{\inf}$ is a non-decreasing function in its second argument and $\mathcal{K}_{\inf}\big(\nu,E(\nu)\big)=0$ for all distributions ν. Therefore,

$$\begin{aligned}\mathbb{E}\big[N_T(a)\big] \leqslant 1 &+ \sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{\mu^\star-\varepsilon<\widehat{\mu}_{a,N_t(a)} \ \text{ and } \ A_{t+1}=a\Big\} + \sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{\mu^\star-\varepsilon>B^+_{a^\star,t}\Big\}\\ &+ \sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{N_t(a)\ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,N_t(a)},\ \mu^\star-\varepsilon\big)\leqslant f(t) \ \text{ and } \ A_{t+1}=a\Big\};\end{aligned}$$

now, the two sums with the events "and $A_{t+1}=a$" can be rewritten by using the stopping times $\tau_{a,k}$ introduced in the proof of Theorem 2.1; more precisely, by mimicking the transformations performed in its step 3, we get the simpler bound

$$\begin{aligned}\mathbb{E}\big[N_T(a)\big] \leqslant 1 &+ \sum_{k=2}^{T-|\mathcal{A}|+1}\mathbb{P}\Big\{\mu^\star-\varepsilon<\widetilde{\mu}_{a,k-1}\Big\} + \sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{\mu^\star-\varepsilon>B^+_{a^\star,t}\Big\}\\ &+ \sum_{k=2}^{T-|\mathcal{A}|+1}\mathbb{P}\Big\{(k-1)\ \mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k-1},\ \mu^\star-\varepsilon\big)\leqslant f(t)\Big\}, \quad (2.6)\end{aligned}$$

where the $\widetilde{\nu}_{a,s}$ and $\widetilde{\mu}_{a,s}$ are respectively the empirical distributions and empirical expectations computed on the first s elements of the sequence of the random variables $\widetilde{X}_{a,j}=Y_{\tau_{a,j}}$, which are i.i.d. according to ν_a.

Step 1: The first sum in (2.6) is bounded by resorting to Hoeffding's inequality, whose application is legitimate since $\mu^\star-\mu_a-\varepsilon>0$;

$$\begin{aligned}\sum_{k=2}^{T-|\mathcal{A}|+1}\mathbb{P}\Big\{\mu^\star-\varepsilon<\widetilde{\mu}_{a,k-1}\Big\} &= \sum_{k=1}^{T-|\mathcal{A}|}\mathbb{P}\Big\{\mu^\star-\mu_a-\varepsilon<\widetilde{\mu}_{a,k}-\mu_a\Big\}\\ &\leqslant \sum_{k=1}^{T-|\mathcal{A}|}e^{-2k(\mu^\star-\mu_a-\varepsilon)^2} \leqslant \frac{1}{1-e^{-2(\mu^\star-\mu_a-\varepsilon)^2}} \leqslant \frac{1}{(\mu^\star-\mu_a-\varepsilon)^2},\end{aligned}$$

where we used for the last inequality the general upper bounds provided at the beginning of step 5 in the proof of Theorem 2.1.

Step 2: The second sum in (2.6) is bounded by first using the definition of $B^+_{a^\star,t}$, then, decomposing the event depending on the values taken by $N_t(a^\star)$; and finally using the fact that on $\{N_t(a^\star)=k\}$, we have the rewriting $\widehat{\nu}_{a,N_t(a)}=\widetilde{\nu}_{a,k}$ and $\widehat{\mu}_{a,N_t(a)}=\widetilde{\mu}_{a,k}$; more precisely,

$$\begin{aligned}\sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{\mu^\star-\varepsilon>B^+_{a^\star,t}\Big\} &\leqslant \sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{N_t(a^\star)\,\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,N_t(a^\star)},\,\mu^\star-\varepsilon\big)>f(t)\Big\}\\ &= \sum_{t=|\mathcal{A}|}^{T-1}\sum_{k=1}^{t}\mathbb{P}\Big\{N_t(a^\star)=k\ \text{ and }\ k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)>f(t)\Big\}\\ &\leqslant \sum_{k=1}^{T}\sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)>f(t)\Big\}.\end{aligned}$$

Since $f=\log$ is increasing, we can rewrite the bound, using a Fubini-Tonelli argument, as

$$\begin{aligned}\sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{\mu^\star-\varepsilon>B^+_{a^\star,t}\Big\} &\leqslant \sum_{k=1}^{T}\sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{f^{-1}\Big(k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)\Big)>t\Big\}\\ &\leqslant \sum_{k=1}^{T}\mathbb{E}\Big[f^{-1}\Big(k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)\Big)\,\mathbb{I}_{\left\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon)>0\right\}}\Big].\end{aligned}$$

Now, Honda and Takemura (2010a, Lemma 13) indicates that, since $\mu^\star-\varepsilon\in[0,1)$,

$$\sup_{\nu\in\mathcal{P}_F([0,1])}\mathcal{K}_{\inf}\big(\nu,\mu^\star-\varepsilon\big)\leqslant\log\big(1/(1-\mu^\star+\varepsilon)\big)\stackrel{\text{def}}{=}K_{\max};$$

we define $Q=K_{\max}/\varepsilon^2$ and introduce the following sets $(V_q)_{1\leqslant q\leqslant Q}$:

$$V_q=\Big\{\nu\in\mathcal{P}_F\big([0,1]\big):\quad(q-1)\varepsilon^2<\mathcal{K}_{\inf}\big(\nu,\mu^*-\varepsilon\big)\leqslant q\varepsilon^2\Big\}.$$

A peeling argument (and by using that $f^{-1}=\exp$ is increasing as well) entails, for all $k\geqslant 1$,

$$\begin{aligned}&\mathbb{E}\Big[f^{-1}\Big(k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)\Big)\,\mathbb{I}_{\left\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon)>0\right\}}\Big] &(2.7)\\ &=\sum_{q=1}^{Q}\mathbb{E}\Big[f^{-1}\Big(k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)\Big)\,\mathbb{I}_{\left\{\widetilde{\nu}_{a^\star,k}\in V_q\right\}}\Big]\\ &\leqslant\sum_{q=1}^{Q}\mathbb{P}\big\{\widetilde{\nu}_{a^\star,k}\in V_q\big\}\,f^{-1}(kq\varepsilon^2)\leqslant\sum_{q=1}^{Q}\mathbb{P}\Big\{\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)>(q-1)\varepsilon^2\Big\}\,f^{-1}(kq\varepsilon^2) &(2.8)\end{aligned}$$

where we used the definition of V_q to obtain each of the two inequalities. Now, by Lemma 2.5, when $E\big(\widetilde{\nu}_{a^\star,k}\big) < \mu^\star - \varepsilon$, which is satisfied whenever $\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\, \mu^\star - \varepsilon\big) > 0$, we have

$$\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\, \mu^\star - \varepsilon\big) \leqslant \mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\, \mu^\star\big) - 2\varepsilon^2 \leqslant \mathcal{K}\big(\widetilde{\nu}_{a^\star,k},\, \nu^\star\big) - 2\varepsilon^2\,,$$

where the last inequality is by mere definition of $\mathcal{K}_{\inf}$. Therefore,

$$\mathbb{P}\Big\{\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\, \mu^\star - \varepsilon\big) > (q-1)\varepsilon^2\Big\} \leqslant \mathbb{P}\Big\{\mathcal{K}\big(\widetilde{\nu}_{a^\star,k},\, \nu^\star\big) > (q+1)\varepsilon^2\Big\}\,.$$

We note that for all $k \geqslant 1$, $\quad \mathbb{P}\Big\{\mathcal{K}\big(\widetilde{\nu}_{a^\star,k},\, \nu^\star\big) > (q+1)\varepsilon^2\Big\} \leqslant (k+1)^{|\mathcal{S}^\star|}\, e^{-k(q+1)\varepsilon^2}\,,$
where we recall that $\mathcal{S}^\star$ denotes the finite support of $\nu^\star$ and where we applied Corollary 2.1 of the appendix. Now, (2.8) then yields, via the choice $f = \log$ and thus $f^{-1} = \exp$, that

$$\mathbb{E}\Big[f^{-1}\Big(k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\, \mu^\star - \varepsilon\big)\Big)\, \mathbb{I}_{\left\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^\star,k},\, \mu^\star - \varepsilon) > 0\right\}}\Big] \leqslant \underbrace{\sum_{q=1}^{Q} (k+1)^{|\mathcal{S}^\star|}\, e^{-k(q+1)\varepsilon^2} e^{kq\varepsilon^2}}_{=Q\,(k+1)^{|\mathcal{S}^\star|}\, e^{-k\varepsilon^2}}\,.$$

Substituting the value of Q, we therefore have proved that

$$\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big\{\mu^\star - \varepsilon > B^+_{a^\star,t}\Big\} \leqslant \frac{1}{\varepsilon^2} \log\left(\frac{1}{1-\mu^* + \varepsilon}\right) \sum_{k=1}^{T} (k+1)^{|\mathcal{S}^\star|}\, e^{-k\varepsilon^2}.$$

Step 3: The third sum in (2.6) is first upper bounded by Lemma 2.5, which states that

$$\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k-1},\, \mu^\star\big) - \varepsilon/(1-\mu^\star) \leqslant \mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k-1},\, \mu^\star - \varepsilon\big)$$

for all $k \geqslant 1$, and by using $f(t) \leqslant f(T)$; this gives

$$\begin{aligned}
&\sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\Big\{k\; \mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k},\, \mu^\star - \varepsilon\big) \leqslant f(t)\Big\} \\
&\qquad\qquad \leqslant \sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\left\{k\; \mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k},\, \mu^\star\big) \leqslant f(T) + \frac{k\,\varepsilon}{1-\mu^\star}\right\} = \sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\Big\{\widetilde{\nu}_{a,k} \in \overline{\mathcal{C}}_{\mu^\star,\gamma_k}\Big\}\,,
\end{aligned}$$

where $\gamma_k = f(T)/k + \varepsilon/(1-\mu^\star)$ and where the set $\overline{\mathcal{C}}_{\mu^\star,\gamma_k}$ was defined in Section 4.1. For all $\gamma > 0$, we then introduce

$$\theta_a(\gamma) = \inf\Big\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \in \mathcal{C}_{\mu^\star,\gamma}\Big\} = \inf\Big\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \in \overline{\mathcal{C}}_{\mu^\star,\gamma}\Big\}\,,$$

(where the second equality follows from the lower semi-continuity of $\mathcal{K}$) and aim at bounding $\mathbb{P}\Big\{\widetilde{\nu}_{a,k} \in \overline{\mathcal{C}}_{\mu^\star,\gamma}\Big\}$.

As shown in Section 4.1, the set $\mathcal{C}_{\mu^\star,\gamma}$ is a non-empty open convex set. If we prove that $\theta_a(\gamma)$ is finite for all $\gamma > 0$, then all the conditions will be required to apply Lemma 2.1 and get the upper bound

$$\sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\Big\{\widetilde{\nu}_{a,k} \in \overline{\mathcal{C}}_{\mu^\star,\gamma_k}\Big\} \leqslant \sum_{k=1}^{T-|\mathcal{A}|} e^{-k\,\theta_a(\gamma_k)}\,.$$

To that end, we use the fact that ν_a is finitely supported. Now, either the probability of interest is null and we are done; or, it is not null, which implies that there exists a possible value of $\widetilde{\nu}_{a,k}$ that is in $\overline{\mathcal{C}}_{\mu^\star,\gamma}$; since this value is a distribution with a support included in the one of ν_a, it is absolutely continuous with respect to ν_a and hence, the Kullback-Leibler divergence between this value and ν_a is finite; in particular, $\theta_a(\gamma)$ is finite.

Finally, we bound the $\theta_a(\gamma_k)$ for values of k larger than $\displaystyle k_0 = \left\lceil \frac{(1+c_a)\,f(T)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)} \right\rceil$; we have that for all $k \geqslant k_0$, in view of the bound put on ε,

$$\gamma_k \leqslant \gamma_{k_0} = \frac{f(T)}{k_0} + \frac{\varepsilon}{1-\mu^\star} < \frac{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}{1+c_a} + \frac{c_a/2}{1+c_a}\,\mathcal{K}_{\inf}(\nu_a,\mu^\star) = \frac{1+c_a/2}{1+c_a}\,\mathcal{K}_{\inf}(\nu_a,\mu^\star)\,. \quad (2.9)$$

Since θ_a is non increasing, we have

$$\sum_{k=1}^{T-|\mathcal{A}|} e^{-k\,\theta_a(\gamma_k)} \leqslant k_0 - 1 + \sum_{k=k_0}^{T-|\mathcal{A}|} e^{-k\,\theta_a(\gamma_{k_0})} \leqslant k_0 - 1 + \frac{1}{1-e^{-\Theta_a(c_a,\varepsilon)}}\,,$$

provided that the quantity $\Theta_a(c_a,\varepsilon) = \theta_a\big(\gamma_{k_0}\big)$ is positive, which we prove now.

Indeed for all $\nu' \in \mathcal{C}_{\mu^\star,\gamma_{k_0}}$, we have by definition and by (2.9) that

$$\mathcal{K}_{\inf}(\nu',\mu^\star) - \mathcal{K}_{\inf}(\nu_a,\mu^\star) < \gamma_{k_0} - \mathcal{K}_{\inf}(\nu_a,\mu^\star) < -\big((c_a/2)/(1+c_a)\big)\mathcal{K}_{\inf}(\nu_a,\mu^\star)\,.$$

Now, in the case where $\mathbb{E}_{\nu_a}\big[(1-\mu^\star)/(1-X)\big] > 1$, we have, first by application of Pinsker's inequality and then by Lemma 2.4, that

$$\mathcal{K}\big(\nu',\nu_a\big) \geqslant \frac{\|\nu'-\nu_a\|_1^2}{2} \geqslant \frac{1}{2\,M^2_{\nu_a,\mu^\star}}\big(\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \mathcal{K}_{\inf}(\nu',\mu^\star)\big)^2 > \frac{c_a^2\,\big(\mathcal{K}_{\inf}(\nu_a,\mu^\star)\big)^2}{8\,(1+c_a)^2\,M^2_{\nu_a,\mu^\star}}\,;$$

since, again by Pinsker's inequality, $\mathcal{K}_{\inf}(\nu_a,\mu^\star) \geqslant (\mu_a-\mu^\star)^2/2 > 0$, we have exhibited a lower bound independent of ν' in this case. In the case where $\mathbb{E}_{\nu_a}\big[(1-\mu^\star)/(1-X)\big] \leqslant 1$, we apply the second part of Lemma 2.4, with $\alpha_a = (c_a/2)/(1+c_a)$, and get

$$\mathcal{K}\big(\nu',\nu_a\big) \geqslant \frac{\|\nu'-\nu_a\|_1^2}{2} \geqslant \frac{1}{2}\left(\frac{1-\mu^\star}{(2/\alpha_a)\,\big((2/\alpha_a)-1\big)}\right)^2 > 0\,.$$

Thus, in both cases we found a positive lower bound independent of ν', so that the infimum over $\nu' \in \mathcal{C}_{\mu^\star,\gamma_{k_0}}$ of the quantities $\mathcal{K}_{\inf}(\nu',\mu^\star)$, which precisely equals $\theta_a\big(\gamma_{k_0}\big)$, is also positive. This concludes the proof. □

Conclusion. We provided a finite-time analysis of the (asymptotically optimal) $\mathcal{K}_{\inf}$–strategy in the case of finitely supported distributions. One could think that the extension to the case of general distributions is straightforward. However this extension appears somewhat difficult (at least when using the current definition of $\mathcal{K}_{\inf}$) for the following reasons: (1) Step 2 in the proof uses the method of types, that would require some extension of Sanov's non-asymptotic Theorem to this case. (2) Step 3 requires to have both $\theta_a(\gamma) < \infty$ for all $\gamma > 0$ and $\theta_a(\gamma) > 0$ for $\gamma < \mathcal{K}_{\inf}(\nu_a, \mu^\star)$, which does not seem to be always the case for general distributions. Exploring other directions for such extensions is left for future work; for instance, histogram-based approximations of general distributions could be considered.

5 Technical details

A conference version of this chapter was published in the *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory* (COLT'11); this appendix details some material which was alluded at in this conference version but could not be published therein because of the page limit.

5.1 Proof of Lemma 2.2

We only provide it for the convenience of the readers since it is similar to the one presented in Garivier and Leonardi (2010, Proposition 2) or in Garivier and Cappé (2011); it was however somewhat simplified by noting that the proof technique used leads to a maximal inequality, as stated in Lemma 2.2, and not only to an inequality for a self-normalized average, as stated in the original reference.

Proof: The result is straightforward in the cases $p = 0$ or $p = 1$, since then, $\widehat{p}_s = p$ almost surely; in the rest of the proof, we therefore only consider the case where $p \in (0, 1)$.

It suffices to show the first bound stated in the lemma, since the second one follows by a decomposition of the probability space according to the values of N_t. Actually, we will show

$$\mathbb{P}_p\left(\bigcup_{s=1}^{t}\left\{ s\ \mathcal{K}\Big(\beta\big(\widehat{p}_s\big),\, \beta(p)\Big) \geqslant \varepsilon \ \text{ and } \ \widehat{p}_s > p \right\}\right) \leqslant e\,\big\lceil \varepsilon \log t \big\rceil\, e^{-\varepsilon}\,,$$

and the desired result will follow by symmetry and a union bound.

Step 1: A martingale. For all $\lambda > 0$, we consider the log-Laplace transform

$$\psi_p(\lambda) = \log \mathbb{E}_p\big[e^{\lambda X_1}\big] = \log\big((1-p) + p\, e^{\lambda}\big)\,,$$

with which we define the martingale

$$W_s(\lambda) = \exp\big(\lambda(X_1 + \ldots + X_s) - s\,\psi_p(\lambda)\big)\,.$$

Step 2: A peeling argument. We introduce $t_0 = 1$ and $t_k = \lfloor \gamma^k \rfloor$, for some $\gamma > 1$ that will be defined by the analysis. We also denote by $K = \lceil (\log t)/(\log \gamma) \rceil$ an upper bound on the number of elements in the peeling.

We also note that by continuity of the Kullback-Leibler divergence in the case of Bernoulli distributions, for all $\varepsilon > 0$, there exists a unique element $p_\varepsilon \in (p, 1)$ such that $\mathcal{K}\big(\beta(q_\varepsilon),\, \beta(p)\big) = \varepsilon$; this element satisfies that

$$\mathcal{K}\big(\beta(q),\, \beta(p)\big) \geqslant \varepsilon \;\text{ and }\; q \geqslant p \qquad \text{entails} \qquad q \geqslant p_\varepsilon \,.$$

Denoting by $\varepsilon_k = \varepsilon / t_k$, a union bound using the described peeling then yields

$$\begin{aligned}
&\mathbb{P}_p\left(\bigcup_{s=1}^{t} \Big\{ s\; \mathcal{K}\Big(\beta\big(\widehat{p}_s\big),\, \beta(p)\Big) \geqslant \varepsilon \;\text{ and }\; \widehat{p}_s > p \Big\} \right) \\
&\quad \leqslant \sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ s\; \mathcal{K}\Big(\beta\big(\widehat{p}_s\big),\, \beta(p)\Big) \geqslant \varepsilon \;\text{ and }\; \widehat{p}_s > p \Big\} \right) \\
&\quad \leqslant \sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ \mathcal{K}\Big(\beta\big(\widehat{p}_s\big),\, \beta(p)\Big) \geqslant \frac{\varepsilon}{t_k} \;\text{ and }\; \widehat{p}_s > p \Big\} \right) \\
&\quad = \sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ \widehat{p}_s \geqslant p_{\varepsilon_k} \Big\} \right) \;=\; \sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ X_1 + \ldots + X_s - s\, p_{\varepsilon_k} \geqslant 0 \Big\} \right)
\end{aligned}$$

Now, the variational formula for Kullback-Leibler divergences shows that for all k, there exists a λ_k such that

$$\varepsilon_k = \mathcal{K}\big(\beta(p_{\varepsilon_k}),\, \beta(p)\big) = \lambda_k\, p_{\varepsilon_k} - \psi_p(\lambda_k)\,;$$

actually, a straightforward calculation shows that $\lambda_k = \log\big(p_{\varepsilon_k}(1-p)\big) - \log\big(p(1-p_{\varepsilon_k})\big) > 0$ is a suitable value. Thus,

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ X_1 + \ldots + X_s - s\, p_{\varepsilon_k} \geqslant 0 \Big\} \right) \\
&\quad = \sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ \exp\big(\lambda_k (X_1 + \ldots + X_s) - \lambda_k s\, p_{\varepsilon_k}\big) \geqslant 1 \Big\} \right) \\
&\quad = \sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ \exp\big(\lambda_k (X_1 + \ldots + X_s) - s\, \psi_p(\lambda_k)\big) \geqslant e^{s\,\varepsilon_k} \Big\} \right) \\
&\quad \leqslant \sum_{k=1}^{K} \mathbb{P}_p\left(\bigcup_{s=t_{k-1}}^{t_k} \Big\{ W_s(\lambda_k) \geqslant e^{t_{k-1}\varepsilon_k} \Big\} \right) \\
&\quad \leqslant \sum_{k=1}^{K} e^{-t_{k-1}\,\varepsilon_k} = K e^{-\varepsilon/\gamma}\,,
\end{aligned}$$

where in the last step, we resorted to Doob's maximal inequality.

Step 3: Choosing γ. The obtained bound equals, by substituting the value of K and by choosing $\gamma = \varepsilon/(\varepsilon - 1)$,

$$Ke^{-\varepsilon/\gamma} = \left\lceil (\log t)/(\log \gamma) \right\rceil e^{-\varepsilon+1} = \left\lceil \frac{\log t}{\log\bigl(\varepsilon/(\varepsilon-1)\bigr)} \right\rceil e^{-\varepsilon+1} ;$$

the proof is concluded by noting that $\varepsilon > 1 \longmapsto \log\bigl(\varepsilon/(\varepsilon-1)\bigr) - 1/\varepsilon$ is decreasing (its derivative is negative), with limit 0 at $+\infty$. □

5.2 Details of the adaptation leading to Lemma 2.1

The exact statement of Dinwoodie (1992, Theorem 2.1 and comments on page 372) is the following.

Lemma 2.7 *[Non-asymptotic Sanov's lemma] Let $\mathcal{C}$ be an open convex subset of $\mathcal{P}(\mathcal{X})$ such that*

$$\Lambda(\mathcal{C}) = \inf_{\kappa \in \mathcal{C}} \mathcal{K}(\kappa, \nu) < \infty .$$

Then, for all $t \geqslant 1$,

$$\mathbb{P}_\nu\bigl\{\widehat{\nu}_t \in \mathcal{C}\bigr\} \leqslant e^{-t\Lambda(\overline{\mathcal{C}})} .$$

We show how it entails Lemma 2.1. Let $\mathcal{C}$ be an open convex subset of $\mathcal{P}(\mathcal{X})$ and let $\overline{\mathcal{C}}$ be its closure. We denote by

$$\mathcal{C}_\delta = \bigl\{\nu \in \mathcal{C} : \quad d(\nu, \mathcal{C}) < \delta\bigr\}$$

the δ–open neighborhood of $\mathcal{C}$, we have $\overline{\mathcal{C}} \subseteq \mathcal{C}_\delta$ for all $\delta > 0$. Therefore, by the lemma above, since $\Lambda(\mathcal{C}_\delta) \leqslant \Lambda(\mathcal{C}) < \infty$,

$$\mathbb{P}_\nu\bigl\{\widehat{\nu}_t \in \overline{\mathcal{C}}\bigr\} \leqslant \mathbb{P}_\nu\bigl\{\widehat{\nu}_t \in \mathcal{C}_\delta\bigr\} \leqslant e^{-t\Lambda(\mathcal{C}_\delta)} .$$

We pick for each integer $n \geqslant 1$ an element κ_n such that $\Lambda\bigl(\mathcal{C}_{1/n}\bigr) = \mathcal{K}(\kappa_n, \nu) - 1/n$; by Dinwoodie (1992, proof of Proposition 1.1), the sequence of the κ_n admits a converging subsequence $\kappa_{\varphi(n)}$, whose limit point κ_∞ belongs to $\overline{\mathcal{C}}$ and which satisfies

$$\mathcal{K}(\kappa_\infty, \nu) \leqslant \liminf_{n\to\infty} \mathcal{K}(\kappa_n, \nu) = \liminf_{\delta \to 0} \Lambda\bigl(\mathcal{C}_\delta\bigr) .$$

Therefore, by taking limits in the above inequality, we have proved the desired inequality,

$$\mathbb{P}_\nu\bigl\{\widehat{\nu}_t \in \overline{\mathcal{C}}\bigr\} \leqslant e^{-t\mathcal{K}(\kappa_\infty,\nu)} \leqslant e^{-t\Lambda(\overline{\mathcal{C}})} .$$

5.3 Useful properties of $\mathcal{K}_{\inf}$ and its level sets

Proof of Lemma 2.4: We resort to the formulation of $\mathcal{K}_{\inf}$ in terms of a convex optimization problem as introduced in Honda and Takemura (2010b); more precisely, it is shown therein that

$$\mathcal{K}_{\inf}(\nu,\mu) = \max\left\{ \mathbb{E}_\nu\Big[\log\big(1+\lambda(\mu-X)\big)\Big] : \quad \lambda \in \big[0,\, 1/(1-\mu)\big] \right\} \tag{2.10}$$

(where X denotes a random variable distributed according to ν), as well as the following alternative. The optimal value λ_ν of the parameter λ indexing the set is equal to $1/(1-\mu)$ if and only if $\mathbb{E}_\nu\big[(1-\mu)/(1-X)\big] \leqslant 1$, and lies in $\big[0,\, 1/(1-\mu)\big)$ if $\mathbb{E}_\nu\big[(1-\mu)/(1-X)\big] > 1$.

For all $\lambda \in \big[0,\, 1/(1-\mu)\big]$, we now introduce the function

$$\varphi_\lambda : x \in [0,1] \;\longmapsto\; \log\big(1+\lambda(\mu-x)\big)\,,$$

which is always continuous on $[0,1)$; we note also that it is continuous and finite at $x=1$ when $\lambda < 1/(1-\mu)$. In the latter case, φ_λ is bounded; since it is decreasing, it is easy to get a uniform bound: for all x,

$$\big|\varphi_\lambda(x)\big| \leqslant \big|\varphi_\lambda(0)\big| + \big|\varphi_\lambda(1)\big| = \log\frac{1+\lambda\mu}{1+\lambda(\mu-1)} \stackrel{\text{def}}{=} M_\lambda\,.$$

It then follows that for all $\lambda \in \big[0,\, 1/(1-\mu)\big)$,

$$\mathbb{E}_\nu\big[\varphi_\lambda(X)\big] - \mathbb{E}_{\nu'}\big[\varphi_\lambda(X)\big] \leqslant M_\lambda \left\|\nu-\nu'\right\|_1 . \tag{2.11}$$

In the case when $\lambda_\nu < 1/(1-\mu)$, we have from the variational formulation (2.10) that

$$\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) \leqslant \mathbb{E}_\nu\big[\varphi_{\lambda_\nu}(X)\big] - \mathbb{E}_{\nu'}\big[\varphi_{\lambda_\nu}(X)\big] \leqslant M_{\lambda_\nu} \left\|\nu-\nu'\right\|_1 .$$

Thus, the constant $M_{\nu,\mu}$ in the statement of the lemma corresponds to our quantity M_{λ_ν} in this case.

We now consider the case where $\lambda_\nu = 1/(1-\mu)$. By (2.11) and variational formulation (2.10), we have that for all $\lambda \in \big[0,\, 1/(1-\mu)\big)$,

$$\begin{aligned}\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) &\leqslant \mathcal{K}_{\inf}(\nu,\mu) - \mathbb{E}_{\nu'}\big[\varphi_\lambda(X)\big] \\ &= \Big(\mathcal{K}_{\inf}(\nu,\mu) - \mathbb{E}_\nu\big[\varphi_\lambda(X)\big]\Big) + \Big(\mathbb{E}_\nu\big[\varphi_\lambda(X)\big] - \mathbb{E}_{\nu'}\big[\varphi_\lambda(X)\big]\Big)\,.\end{aligned}$$

The second difference is bounded according to (2.11); the first difference is bounded by concavity of $\lambda < 1/(1-\mu) \;\mapsto\; \varphi_\lambda(x)$, for all x:

$$\begin{aligned}\mathbb{E}_\nu\big[\varphi_\lambda(X)\big] &\geqslant \big(1-\lambda(1-\mu)\big)\,\mathbb{E}_\nu\big[\varphi_0(X)\big] + \lambda(1-\mu)\,\mathbb{E}_\nu\big[\varphi_0(X)\big] \\ &= \lambda(1-\mu)\,\mathbb{E}_\nu\big[\varphi_{1/(1-\mu)}(X)\big] = \lambda(1-\mu)\,\mathcal{K}_{\inf}(\nu,\mu)\,,\end{aligned}$$

since φ_0 is the null function and $\lambda_\nu = 1/(1-\mu)$. Putting all pieces together, we have proved that for all $\lambda \in \big[0,\, 1/(1-\mu)\big)$,

$$\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) \leqslant \big(1-\lambda(1-\mu)\big)\,\mathcal{K}_{\inf}(\nu,\mu) + M_\lambda \,\|\nu-\nu'\|_1 \,. \tag{2.12}$$

We recall that by assumption, $\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) \geqslant \alpha\,\mathcal{K}_{\inf}(\nu,\mu)$ with $\alpha \in (0,1)$, so that the choice $\lambda = (1-\alpha/2)/(1-\mu)$, which indeed lies in $\big(0,\, 1/(1-\mu)\big)$, is such that

$$M_\lambda = \log\left(1 + \frac{\lambda}{1+\lambda(\mu-1)}\right) = \log\left(1+\frac{\lambda}{\alpha/2}\right) \leqslant \frac{2\lambda}{\alpha}\,,$$

so that (2.12) entails

$$\alpha\,\mathcal{K}_{\inf}(\nu,\mu) \leqslant \frac{\alpha}{2}\,\mathcal{K}_{\inf}(\nu,\mu) + \frac{2\lambda}{\alpha}\,\|\nu-\nu'\|_1\,,$$

and finally

$$\|\nu-\nu'\|_1 \geqslant \frac{\alpha^2}{4\lambda} = \frac{\alpha^2(1-\mu)}{1-\alpha/2} = \frac{1-\mu}{(2/\alpha)\,\big((2/\alpha)-1\big)}\,;$$

which concludes the proof. □

Proof of Lemma 2.5: In Honda and Takemura (2010b) it is shown that in this case, $\mathcal{K}_{\inf}(\nu,\mu)$ is differentiable in $\mu \in (E(\nu),1)$ with

$$\frac{1}{1-\mu} \geqslant \frac{\partial}{\partial\mu}\mathcal{K}_{\inf}(\nu,\mu) \geqslant \frac{\mu-E(\nu)}{\mu(1-\mu)}. \tag{2.13}$$

We apply this result to the rewriting

$$\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) = \int_{\mu-\varepsilon}^{\mu} \frac{\partial}{\partial\mu}\mathcal{K}_{\inf}(\nu,u)\,\mathrm{d}u\,,$$

which already gives one part of the bound. For the lower bound, we note that by assumption $-E(\nu) > -(\mu-\varepsilon)$ and that $u(1-u) \leqslant 1/4$ (since we consider distributions with support included in $[0,1]$); so that, for all u,

$$\frac{u-E(\nu)}{u(1-u)} \geqslant 4\big(u-(\mu-\varepsilon)\big)\,.$$

Integrating the bound concludes the main part of the proof.

Now, to see that the first inequality in the statement is always valid, we need to consider the case when $E(\nu) \geqslant \mu$, for which the statement is trivial since then $\mathcal{K}_{\inf}(\nu,\mu) = 0$, and the case when $\mu > E(\nu) \geqslant \mu-\varepsilon$. But in the latter case, it is shown in Honda and Takemura (2010b, Lemma 6, case 2) that

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

which concludes the proof. □

Proof of Lemma 2.6: First, $\mathcal{C}_\mu(\gamma)$ is non empty as it always contains δ_μ, the Dirac mass on μ.

The fact that $\mathcal{C}_\mu(\gamma)$ is convex follows from the convexity of $\mathcal{K}$ in the pair of probability distributions that it takes as an argument. Indeed, for all $\alpha \in [0,1]$, ν', $\nu'' \in \mathcal{C}_\mu(\gamma)$, denoting by ν'_μ, ν''_μ some distributions such that the defining conditions in $\mathcal{C}_\mu(\gamma)$ are satisfied, we have that

$$E\big(\alpha\nu'_\mu + (1-\alpha)\nu''_\mu\big) > \mu$$

and

$$\mathcal{K}\big(\alpha\nu' + (1-\alpha)\nu'',\, \alpha\nu'_\mu + (1-\alpha)\nu''_\mu\big) \leqslant \alpha\,\mathcal{K}\big(\nu', \nu'_\mu\big) + (1-\alpha)\,\mathcal{K}\big(\nu'', \nu''_\mu\big) < \gamma\,.$$

We prove that $\mathcal{C}_\mu(\gamma)$ is an open set. With each $\nu' \in \mathcal{C}_\mu(\gamma)$, we associate a distribution ν'_μ satisfying the defining constraints in $\mathcal{C}_\mu(\gamma)$; by choosing

$$\alpha = \frac{1 - \mu/E\big(\nu'_\mu\big)}{2} \in (0,\, 1/2),$$

we have that the open set formed by the

$$(1-\alpha)\,\nu' + \alpha\,\nu'', \qquad \nu'' \in \mathrm{B}(\nu', 1)$$

is contained in $\mathcal{C}_{\mu,\gamma}$, where $\mathrm{B}(\nu', 1)$ denotes the ball with center ν' and radius 1 in the norm $\|\cdot\|$ over $\mathcal{P}(\mathcal{X})$. Indeed, we have on the one hand,

$$E\big((1-\alpha)\,\nu'_\mu + \alpha\,\nu''\big) \geqslant (1-\alpha)\,E\big(\nu'_\mu\big) \geqslant \left(1 - \frac{1 - \mu/E\big(\nu'_\mu\big)}{2}\right) E\big(\nu'_\mu\big) = \frac{E\big(\nu'_\mu\big) + \mu}{2} > \mu\,,$$

and on the other hand, by convexity of the Kullback-Leibler divergence,

$$\mathcal{K}\big((1-\alpha)\,\nu' + \alpha\,\nu'',\, (1-\alpha)\,\nu'_\mu + \alpha\,\nu''\big) \leqslant (1-\alpha)\,\mathcal{K}\big(\nu',\, \nu'_\mu\big) < (1-\alpha)\gamma\,.$$

To prove the desired inclusion, we first note that in the case of $\mathcal{P}_F\big([0,1]\big)$, Honda and Takemura (2010b) show that one has the rewriting

$$\mathcal{K}_{\inf}(\nu, \mu) = \min\Big\{\mathcal{K}(\nu, \nu') : \quad \nu' \in \mathcal{P}_F\big([0,1]\big) \;\text{ s.t. }\; E(\nu') \geqslant \mu\Big\}\,;$$

in particular, the infimum is achieved with this new formulation. Hence,

$$\mathcal{C}_{\mu,\gamma} = \Big\{\nu' \in \mathcal{P}_F\big([0,1]\big) : \quad \exists\, \nu'_\mu \in \mathcal{P}_F\big([0,1]\big) \;\text{ s.t. }\; E\big(\nu'_\mu\big) \geqslant \mu \;\text{ and }\; \mathcal{K}\big(\nu', \nu'_\mu\big) < \gamma\Big\}\,.$$

Also, an element of the set of interest is therefore a $\nu' \in \mathcal{P}_F\big([0,1]\big)$ such that $\mathcal{K}_{\inf}(\nu', \mu) \leqslant \gamma$, that is, such that there exists $\nu'_\mu \in \mathcal{P}\big([0,1]\big)$ with $E\big(\nu'_\mu\big) \geqslant \mu$ and $\mathcal{K}\big(\nu', \nu'_\mu\big) \leqslant \gamma$. Now, the distributions

$$\nu'_n = \left(1 - \frac{1}{n}\right)\nu' + \frac{1}{n}\delta_1\,, \qquad \text{thanks to the} \qquad \nu'_{\mu,n} = \left(1 - \frac{1}{n}\right)\nu'_\mu + \frac{1}{n}\delta_1\,,$$

all belong to $\mathcal{C}_\gamma$, as, similarly to the above argument,

$$E\big(\nu'_n\big) \geqslant \mu + \frac{1-\mu}{n} > \mu \qquad \text{and} \qquad \mathcal{K}\big(\nu'_n,\, \nu'_{\mu,n}\big) \leqslant \left(1 - \frac{1}{n}\right)\mathcal{K}\big(\nu', \nu'_\mu\big) < \gamma\,.$$

In addition, we have by construction that the ν'_n converge to ν', hence, $\nu' \in \overline{\mathcal{C}}_\gamma$. □

5.4 The method of types

Let $X_1, X_2, \ldots$ be a sequence of random variables that are i.i.d. according to a distribution denoted by ν. In this subsection, we will index all probabilities and expectations by ν.

For all $k \geqslant$, we denote by $\mathcal{E}_k$ the set of possible values (the so-called types) of the empirical distribution

$$\widehat{\nu}_k = \sum_{j=1}^{k} \delta_{X_j} \, .$$

If ν has a finite support denoted by $\mathcal{S}$, then the cardinality $|\mathcal{E}_k|$ of $\mathcal{E}_k$ is bounded by $(k+1)^{|\mathcal{S}|}$.

Lemma 2.8 *In the case where ν has a finite support, for all $k \geqslant 1$ and $\kappa \in \mathcal{E}_k$,*

$$\mathbb{P}_\nu\big\{\widehat{\nu}_k = \kappa\big\} \leqslant e^{-k\,\mathcal{K}(\kappa,\nu)} \, .$$

Corollary 2.1 *In the case where ν has a finite support, for all $k \geqslant 1$, all $\gamma > 0$,*

$$\begin{aligned}\mathbb{P}\Big\{\mathcal{K}\big(\widehat{\nu}_k,\, \nu\big) > \gamma\Big\} &= \sum_{\kappa \in \mathcal{E}_k} \mathbb{I}_{\{\mathcal{K}(\kappa,\nu)>\gamma\}}\, \mathbb{P}_\nu\big\{\widehat{\nu}_k = \kappa\big\} \\ &\leqslant \sum_{\kappa \in \mathcal{E}_k} \mathbb{I}_{\{\mathcal{K}(\kappa,\nu)>\gamma\}}\, e^{-k\,\mathcal{K}(\kappa,\nu)} \leqslant |\mathcal{E}_k|\, e^{-k\gamma} \leqslant (k+1)^{|\mathcal{S}|}\, e^{-k\gamma} \, .\end{aligned}$$

Chapter 3

Bandit Algorithms for Online Learning in Adversarial Lipschitz Environments.

In this chapter, we now leave the stochastic multi-armed bandit problem and turn to the setting of multi-armed bandit with adversarial environment but full information. Since the setting of full information enables to deal with a large set of arms, we drop the assumption that the set of arms $\mathcal{A}$ is finite, and consider it is some subset of $\mathbb{R}^d$, which enables to address the problem of online learning in an adversarial environment. For such large sets, one has to assume some regularity on the reward functions in order to control the regret term. The main difficulty, however, is to derive efficient numerical implementations for such settings, which generally requires to make approximations of a theoretical algorithm.

Here we consider the problem of online learning in an adversarial environment when the reward functions chosen by the adversary are assumed to be Lipschitz. This setting extends previous works on linear (see Dani et al. (2008a), Abernethy et al. (2008b), Cesa-Bianchi and Lugosi (2009), Kakade et al. (2008)) and convex (see Zinkevich (2003), Hazan et al. (2006)) online learning. We provide a class of algorithms with cumulative regret upper bounded by $\tilde{O}(\sqrt{dT\ln(\lambda)})$ where d is the dimension of the search space, T the time horizon, and λ the Lipschitz constant. We discuss the major issue of deriving efficient numerical implementations and makes use of particle methods for this purpose. Applications include online supervised learning problems for both full and partial (bandit) information settings, for a large class of non-linear regressors/classifiers, such as neural networks.

This work has been published in the proceedings of the *21st European Conference on Machine Learning (ECML 2010)*, see Maillard and Munos (2010b) for details.

Contents

Introduction

The adversarial online learning problem is defined as a repeated game between an agent (the learner) and an opponent, where at each round t, simultaneously the agent chooses an action (or decision, or arm, or state) $\theta_t \in \Theta$ (where Θ is a subset of $\mathbb{R}^d$) and the opponent chooses a reward function $f_t : \Theta \mapsto [0,1]$. The agent receives the reward $f_t(\theta_t)$. In this chapter we will consider different assumptions about the amount of information received by the agent at each round. In the *full information* case, the full reward function f_t is revealed to the agent after each round, whereas in the case of *bandit information* only the reward corresponding to its own choice $f_t(\theta_t)$ is provided.

The goal of the agent is to allocate its actions $(\theta_t)_{1 \leqslant t \leqslant T}$ in order to maximize the sum of obtained rewards $F_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} f_t(\theta_t)$ up to time T and its performance is assessed in terms of the best constant strategy $\theta \in \Theta$ on the same reward functions, i.e. $F_T(\theta) \stackrel{\text{def}}{=} \sum_{t=1}^{T} f_t(\theta)$. Defining the **cumulative regret**:

$$R_T(\theta) \stackrel{\text{def}}{=} F_T(\theta) - F_T,$$

with respect to (w.r.t.) a strategy θ, the agent aims at minimizing $R_T(\theta)$ for all $\theta \in \Theta$.

In this work we consider the case when the functions f_t are Lipschitz w.r.t. the decision variable θ (with Lipschitz constant upper bounded by λ).

Previous results. Several works on adversarial online learning include the case of finite action spaces (the so-called *learning from experts* Cesa-Bianchi and Lugosi (2006) and the *multi-armed bandit problem* Auer et al. (1995, 2003)), countably infinite action spaces Poland (2008), and the case of continuous action spaces, where many works have considered strong assumptions on the reward functions, i.e. linearity or convexity.

In the *online linear optimization* (see e.g. Dani et al. (2008a), Abernethy et al. (2008b), Cesa-Bianchi and Lugosi (2009), Kakade et al. (2008) in the adversarial case and Auer (2003), Dani et al. (2008b) in the stochastic case) where the functions f_t are linear, the resulting upper- and lower-bounds on the regret are of order (up to logarithmic factors) $\sqrt{dT}$ in the case of full information and $d^{3/2}\sqrt{T}$ in the case of bandit information Abernethy et al. (2008b) (and in good cases $d\sqrt{T}$ Dani et al. (2008a)). In *online convex optimization* f_t is assumed to be convex Zinkevich (2003) or σ-strongly convex Hazan et al. (2006), and the resulting upper bounds are of order $C\sqrt{T}$ and $C^2\sigma^{-1}\ln(T)$ (where C is a bound on the gradient of

the functions, which implicitly depends on the space dimension). Other extensions have been considered in Bartlett et al. (2007), Shalev-Shwartz (2007), Flaxman et al. (2005) and a minimax lower bound analysis in the full information case in Abernethy et al. (2008a). These results hold in bandit information settings where either the value or the gradient of the function is revealed.

To our knowledge, the weaker Lipschitz assumption that we consider here has not been studied in the adversarial optimization literature. However, in the stochastic bandit setting (where noisy evaluations of a fixed function are revealed), the Lipschitz assumption has been previously considered in Kleinberg et al. (2008), Bubeck et al. (2008), see the discussion in Section 2.3.

Motivations. In many applications (such as the problem of matching ads to web-page contents on the Internet) it is important to be able to consider both large action spaces and general reward functions. The continuous space problem appears naturally in online learning, where a decision point is a classifier in a parametric space of dimension d. Since many non-linear non-convex classifiers/regressors have shown success (such as neural-networks, support vector machines, matching pursuits), we wish to extend the results of online learning to those non-linear non-convex cases. In this work we consider a Lipschitz assumption (illustrated in the case of neural network architectures) which is much weaker than linearity or convexity.

Contribution. We start in Section 1 by describing a general continuous version of the Exponentially Weighted Forecaster and state (Theorem 3.1) an upper bound on the cumulative regret of $O(\sqrt{dT\ln(d\lambda T)})$ under a non-trivial geometrical property of the action space. The algorithm requires, as a sub-routine, being able to sample actions according to continuous distributions, which may be impossible to do perfectly well in general.

To address the issue of sampling, we may use different sampling techniques, such as uniform grids, random or quasi-random grids, or use adaptive methods such as Monte-Carlo Markov chains (MCMC) or Population Monte-Carlo (PMC).

However, since any sampling technique introduces a sampling bias (compared to an ideal sampling from the continuous distribution), this also impacts the resulting performance of the method in terms of regret. This shows a tradeoff between regret and numerical complexity, which is illustrated by numerical experiments in Section 1.3 where PMC techniques are compared to sampling from uniform grids.

Then in Section 2 we describe several applications to learning problems. In the full information setting (when the desired outputs are revealed after each round), the case of regression is described in Section 2.1 and the case of classification in Section 2.2. Then Section 2.3 considers a classification problem in a bandit setting (i.e. when only the information of whether the prediction is correct or not is revealed). In the later case, we show that the expected number of mistakes does not exceed that of the best classifier by more than $O(\sqrt{dTK\ln(d\lambda T)})$, where K is the number of labels. We detail a possible PMC implementation in this case.

We believe that the work reported in this chapter provides arguments that the use of MCMC, PMC, and other adaptive sampling techniques is a promising direction for designing numerically efficient algorithms for online learning in adversarial Lipschitz environments.

1 Adversarial learning with full information

We consider a search space $\Theta \subset \mathbb{R}^d$ equipped with the Lebesgue measure μ. We write $\mu(\Theta) = \int_\Theta 1$. We assume that all reward functions f_t have values in $[0,1]$ and are Lipschitz w.r.t. some norm $||\cdot||$ (e.g. L_1, L_2, or L_∞) with a Lipschitz constant upper bounded by $\lambda > 0$, i.e. for all $t \geqslant 1$ and $\theta_1, \theta_2 \in \Theta$,

$$|f_t(\theta_1) - f_t(\theta_2)| \leqslant \lambda ||\theta_1 - \theta_2||.$$

1.1 The ALF strategy

We consider the natural extension of the EWF (Exponentially Weighted Forecaster) algorithm Littlestone and Warmuth (1989), Cesa-Bianchi et al. (1997), Cesa-Bianchi and Lugosi (2006) to the continuous action setting. Figure 3.1 describes this ALF strategy (for Adversarial Lipschitz Full-information environment).

Initialization: Set $w_1(\theta) = 1$ for all $\theta \in \Theta$.

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

(1) Simultaneously the adversary chooses the reward function $f_t : \Theta \mapsto [0,1]$, and the learner chooses $\theta_t \overset{iid}{\sim} p_t$, where $p_t(\theta) \overset{\text{def}}{=} \dfrac{w_t(\theta)}{\int_\Theta w_t(\theta) d\theta}$,

(2) The learner incurs the reward $f_t(\theta_t)$,

(3) The reward function f_t is revealed to the learner. The weight function w_t is updated as:

$$w_{t+1}(\theta) \overset{\text{def}}{=} w_t(\theta) e^{\eta f_t(\theta)}, \text{ for all } \theta \in \Theta$$

Figure 3.1: Adversarial Lipschitz learning algorithm in a Full-information setting (ALF strategy)

At each time step, the forecaster samples θ_t from a probability distribution $p_t \overset{\text{def}}{=} \frac{w_t}{\int_\Theta w_t}$ with w_t being the weight function defined according to the previously observed reward functions $(f_s)_{s<t}$. The function f_t is then revealed and the weight function is updated. We have $w_{t+1}(\theta) = \exp(\eta F_t(\theta))$, and η is a parameter of the algorithm.

Geometric considerations: The performance of the algorithm depends on the geometry of the space $\Theta \subset \mathbb{R}^d$ (relatively to the chosen norm), and since we want to derive bounds as a function of the dimension d, we now define classes of domains ($(\Theta_d)_{d\geqslant 0}$ indexed by their dimension) with similar geometrical properties.

Definition 3.1 *For the class of domains $(\Theta_d)_{d\geqslant 1}$, we define $\kappa(d) > 1$:*

$$\kappa(d) \stackrel{\text{def}}{=} \sup_{\theta\in\Theta_d, r>0} \frac{\min\big[\mu\big(B(\theta,r)\big), \mu(\Theta_d)\big]}{\mu\big(B(\theta,r)\cap\Theta_d\big)} \tag{3.1}$$

Assumption A1 There exists $\kappa > 0$ such that $\kappa(d) \leqslant \kappa^d$, for all $d \geqslant 1$, and there exists $\kappa' > 0$ and $\alpha \geqslant 0$ such that $\mu(B(\theta,r)\big) \geqslant (r/(\kappa' d^\alpha))^d$ for all $r > 0$, $d \geqslant 1$, and $\theta \in \mathbb{R}^d$.

The first part of this assumption says that $\kappa(d)$ scales at most exponentially with the dimension. This is reasonable if we consider domains with similar geometries (i.e. whenever the "angles" of the domains do not go to zero when the dimension d increases). For example, in the domains $\Theta_d = [0,1]^d$, this assumption holds with $\kappa = 2$ for any usual norm (L_1,L_2 and L_∞). The second part of the assumption about the volume of d-balls is a property of the norms and holds naturally for any usual norm: for example, $\kappa' = 1/2$, $\alpha = 0$ for L_∞, and $\kappa' = \sqrt{\pi}/(\sqrt{2}e)$, $\alpha = 3/2$ for any norm L_p, $p \geqslant 1$, since for L_p norms, $\mu(B(\theta,r)) \geqslant (2r)^d/d!$ and from Stirling formula, $d! \sim \sqrt{2\pi d}(d/e)^d$, thus $\mu(B(\theta,r)) \geqslant \big(r/(\frac{\sqrt{2\pi}}{2e}d^{3/2})\big)^d$.

Remark 2 *Notice that Assumption A1 makes explicit the required geometry of the domain in order to derive tight regret bounds.*

We now provide upper-bounds for the ALF strategy on the worst expected regret (i.e. $\sup_{\theta\in\Theta} \mathbb{E}R_T(\theta)$) and high probability bounds on the worst regret $\sup_{\theta\in\Theta} R_T(\theta)$.

Theorem 3.1 (*Regret bound for the ALF strategy*) *Under Assumption A1, for any $\eta \leqslant 1$, the expected (w.r.t. the internal randomization of the algorithm) cumulative regret of the ALF strategy is bounded as:*

$$\sup_{\theta\in\Theta} \mathbb{E}R_T(\theta) \leqslant T\eta + \frac{1}{\eta}\Big[d\ln(cd^\alpha\eta\lambda T) + \ln(\mu(\Theta))\Big], \tag{3.2}$$

whenever $(d^\alpha\eta\lambda T)^d\mu(\Theta) \geqslant 1$, where $c \stackrel{\text{def}}{=} 2\kappa\max(\kappa', 1)$ is a constant (which depends on the geometry of Θ and the considered norm). Under the same assumptions, with probability $1-\beta$,

$$\sup_{\theta\in\Theta} R_T(\theta) \leqslant T\eta + \frac{1}{\eta}\Big[d\ln(cd^\alpha\eta\lambda T) + \ln(\mu(\Theta))\Big] + \sqrt{2T\ln(\beta^{-1})}. \tag{3.3}$$

We deduce that for the choice $\eta = \big(\frac{d}{T}\ln(cd^\alpha\lambda T)\big)^{1/2}$, when $\eta \leqslant 1$ and assuming $\mu(\Theta) = 1$, we have:

$$\sup_{\theta\in\Theta} \mathbb{E}R_T(\theta) \leqslant 2\sqrt{dT\ln(cd^\alpha\lambda T)},$$

and a similar bound holds in high probability.

The proof is given in Appendix 4. Note that the parameter η of the algorithm depends very mildly on the (unknown) Lipschitz constant λ. Actually even if λ was totally unknown, the choice $\eta = \left(\frac{d}{T}\ln(cd^\alpha T)\right)^{1/2}$ would yield a bound $\sup_{\theta\in\Theta}\mathbb{E}R_T(\theta) = O(\sqrt{dT\ln(dT)}\ln\lambda)$ which is still logarithmic in λ (instead of linear in the case of the discretization) and enables to consider classes of functions for which λ may be large (and unknown).

Anytime algorithm. Like in the discrete version of EWF (see e.g. Auer et al. (2000), Stoltz (2005), Cesa-Bianchi and Lugosi (2006)) this algorithms may easily be extended to an anytime algorithm (i.e. providing similar performance even when the time horizon T is not known in advance) by considering a decreasing coefficient $\eta_t = \left(\frac{d}{2t}\ln(cd^\alpha\lambda t)\right)^{1/2}$ in the definition of the weight function w_t. We refer to Stoltz (2005) for a description of the methodology.

The issue of sampling. In order to implement the ALF strategy detailed in Figure 3.1 one should be able to sample θ_t from the continuous distribution p_t. However it is in general impossible to sample perfectly from arbitrary continuous distributions p_t, thus we need to resort to approximate sampling techniques, such as based on uniform grids, random or quasi-random grids, or adaptive methods such as Monte-Carlo Markov Chain (MCMC) methods or population Monte-Carlo (PMC) methods. If we write p_t^N the distribution from which the samples are actually generated, where N stands for the computational resources (e.g. the number of grid points if we use a grid) used to generate the samples, then the expected regret $\mathbb{E}R_T(\theta)$ will suffer an additional term of at most $\sum_{t=1}^T \left|\int_\Theta p_t f_t - \int_\Theta p_t^N f_t\right|$. This shows a tradeoff between the regret (low when N is large, i.e. p_t^N is close to p_t) and numerical complexity and memory requirement (which scales with N). In the next two sub-sections we discuss sampling techniques based on fixed grids and adaptive PMC methods, respectively.

1.2 Uniform grid over the unit hypercube

A first approach consists in setting a uniform grid (say with N grid points) before the learning starts and consider the naive approximation of p_t by sampling at each round one point of the grid, since in that case the distribution has finite support and the sampling is easy.

Actually, in the case when the domain Θ is the unit hypercube $[0,1]^d$, we can easily do the analysis of an Exponentially Weighted Forecaster (EWF) playing on the grid and shows that the total expected regret is small provided that N is large enough. Indeed, let $\Theta_N \overset{\text{def}}{=} \{\theta_1,\ldots,\theta_N\}$ be a uniform grid of resolution $h>0$, i.e. such that for any $\theta\in\Theta$, $\min_{1\leqslant i\leqslant N}||\theta-\theta_i|| \leqslant h$. This means that at each round t, we select the action $\theta_{I_t}\in\Theta_N$, where $I_t \overset{iid}{\sim} p_t^N$ with p_t^N the distribution on $\{1,\ldots,N\}$ defined by $p_t^N(i) \overset{\text{def}}{=} w_t(i)/\sum_{j=1}^N w_t(j)$, where the weights are defined as $w_t(i) \overset{\text{def}}{=} e^{\eta F_{t-1}(\theta_i)}$ for some appropriate constant $\eta = \sqrt{2\ln N/T}$.

The usual analysis of EWF implies that the regret relatively to any point of the grid is upper bounded as: $\sup_{1\leqslant i\leqslant N}\mathbb{E}R_T(\theta_i) \leqslant \sqrt{2T\ln N}$.

Now, since we consider the unit hypercube $\Theta = [0,1]^d$, and under the assumption that the functions f_t are λ-Lipschitz with respect to L_∞-norm, we have that $F_T(\theta) \leqslant \min_{1\leqslant i\leqslant N} F_T(\theta_i) + \lambda T h$. We deduce that the expected regret relatively to any $\theta \in \Theta$ is bounded as $\sup_{\theta\in\Theta} \mathbb{E}R_T(\theta) \leqslant \sqrt{2T \ln N} + \lambda T h$.

Setting $N = h^{-d}$ with the optimal choice of h in the previous bound (up to a logarithmic term) $h = \frac{1}{\lambda}\sqrt{d/T}$ gives the upper bound on the regret: $\sup_{\theta\in\Theta} \mathbb{E}R_T = O(\sqrt{dT \ln(\lambda\sqrt{T})})$.

However this discretized EWF algorithm suffers from severe limitations from a practical point of view:

1. The choice of the best resolution h of the grid depends crucially on the knowledge of the Lipschitz constant λ and has an important impact on the regret bound. However, usually λ is not known exaclty (but an upper-bound may be available, e.g. in the case of neural networks discussed below). If we choose h irrespective of λ (e.g. $h = \sqrt{d/T}$) then the resulting bound on the regret will be of order $O(\lambda\sqrt{dT})$ which is much worst in terms of λ than its optimal order $\sqrt{\ln \lambda}$.

2. The number of grid points (which determines the memory requirement and the numerical complexity of the EWF algorithm) scales exponentially with the dimension d.

Notice that instead of using a uniform grid, one may resort to the use of random (or quasi-random) grids with a given number of points N, which would scale better in high dimensions. However all those method are non-adaptive in the sense that the position of the grid point do not adapt to the actual reward functions f_t observed through time. We would like to sample points according to an "adaptive discretization" that would allocate more points where the cumulative reward function F_t is high. In the next sub-section we consider the ALF strategy where we use adaptive sampling techniques such as MCMC and PMC which are designed for sampling from (possibly high dimensional) continuous distributions.

1.3 A Population Monte-Carlo sampling technique

The idea of sampling techniques such as Metropolis-Hasting (MH) or other MCMC (Monte-Carlo Markov Chain) methods (see e.g. Gilks et al. (1996), Andrieu et al. (2003)) is to build a Markov chain that has p_t as its equilibrium distribution, and starting from an initial distribution, iterates its transition kernel K times so as to approximate p_t. Note that the rate of convergence of the distribution towards p_t is exponential with K (see e.g. Levin et al. (2008)): $\delta(k) \leqslant (2\varepsilon)^{k/\tau(\varepsilon)}$, where $\delta(k)$ is the total variation distance between p_t and the distribution at step k, and $\tau(\varepsilon) = \min\{k; \delta(k) \leqslant \varepsilon\}$ is the so called mixing time of the Markov Chain ($\varepsilon < 1/2$).

Thus sampling $\theta_t \sim p_t$ only requires being able to compute $w_t(\theta)$ at a finite number of points K (the number of transitions of the corresponding Markov chain needed to approximate the stationary distribution p_t). This is possible whenever the reward functions f_t can

be stored by using a finite amount of information, which is the case in the applications to learning, described in the next section.

However, using MCMC at each time step to sample from a distribution p_t which is similar to the previous one p_{t-1} (since the cumulative functions F_t do not change much from one iteration to the next) is a waste of MC transitions. The exponential decay of $\delta(k)$ depends on the mixing time $\tau(\varepsilon)$ which depends on both the target distribution and the transition kernel, and can be reduced when considering efficient methods based on interacting particles systems. The population Monte-Carlo (PMC) method (see e.g. Douc et al. (2007)) approximates p_t by a population of N particles $(x_{t,k}^{1:N})$ which evolve (during $1 \leqslant k \leqslant K$ rounds) according to a transition/selection scheme:

- At round k, the **transition step** generates a successor population $\widetilde{x}_{t,k}^{1:N} \overset{iid}{\sim} g_{t,k}(x_{t,k-1}^{1:N}, \cdot)$ according to a transition kernel $g_{t,k}(\cdot,\cdot)$. Then likelihood ratios are defined as $w_{t,k}^{1:N} = \frac{p_t(\widetilde{x}_{t,k}^{1:N})}{g(x_{t,k-1}^{1:N}, \widetilde{x}_{t,k}^{1:N})}$,
- The **selection step** resamples N particles $x_{t,k}^i = \widetilde{x}_{t,k}^{I_i}$ for $1 \leqslant i \leqslant N$ where the selection indices $(I_i)_{1\leqslant i\leqslant N}$ are drawn (with replacement) from the set $\{1 \dots N\}$ according to a multinomial distribution with parameters $(w_{t,k}^i)_{1\leqslant i\leqslant N}$

At round K, one particle (out of N) is selected uniformly randomly, which defines the sample $\theta_t \sim p_t^N$ that is returned by the sampling technique. Some properties of this approch is that the proposed sample tends to an unbiased independent sample of p_t (when either N or $K \to \infty$). We do not provide additional implementation *details* about this method here since this is not the goal of this chapter, but we refer the interested reader to Douc et al. (2007) for discussion about the choice of good kernels $g_{t,k}$ and automatic tuning methods of the parameter K and number of particles N. Note that in Douc et al. (2007), the authors prove a Central Limit Theorem showing that the term $\sqrt{N}(\int_\Theta p_t f - \int_\Theta p_t^N f)$ is asymptotically gaussian with explicit variance depending on the previous parameters (that we do not report here for it would require additional specific notations), thus giving the speed of convergence towards 0. We also refer to Del Moral (2004) for known theoretical results of the general PMC theory.

When using this sampling techniques in the ALF strategy, since the distribution p_{t+1} does not differ much from p_t, we can initialize the particles at round $t+1$ with the particles obtained at the previous round t at the last step of the PMC sampling: $x_{t+1,1}^i \overset{\text{def}}{=} x_{t,K}^i$, for $1 \leqslant i \leqslant N$. In the numerical experiments reported in the next sub-section, this enabled to reduce drastically the number of rounds K per time step (less than 5 in all experiments below).

Let us remark that, since the publication of this work, it was proved in Narayanan and Rakhlin (2010) that in the special case when the functions are moreover assumed to be convex, then only one step of MCMC method is sufficient to get a control on the approximation error induced by the sampling scheme. This is due to new results about so-called randomized interior point method by Narayanan (2009).

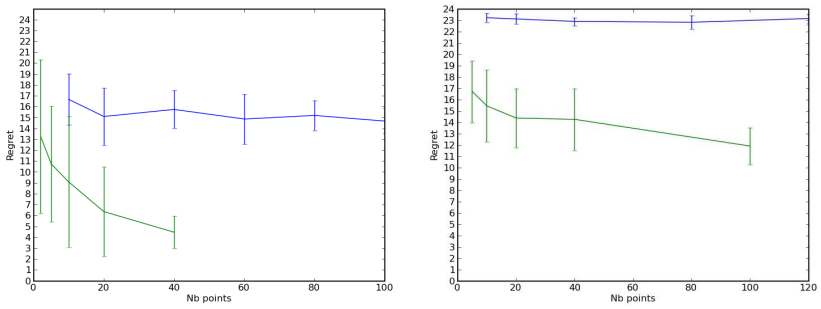


Figure 3.2: Regret as a function of N, for dimensions $d = 2$ (left figure) and 20 (right figure). In both figures, the top curve represents the grid sampling and the bottom curve the PMC sampling

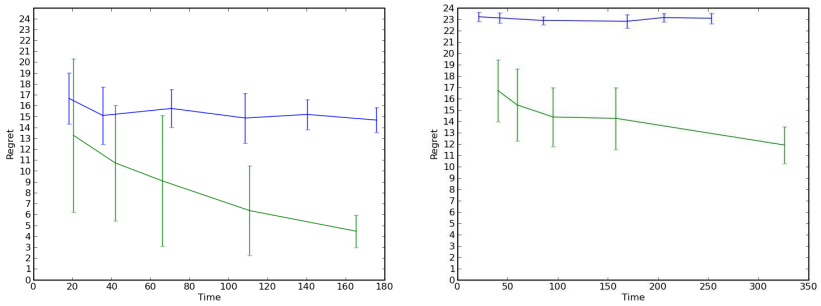


Figure 3.3: Regret as a function of the CPU time used for sampling, for dimensions $d = 2$ (left figure) and 20 (right figure). Again, in both figures, the top curve represents the grid sampling and the bottom curve the PMC sampling.

1.4 Numerical experiments

For illustation, let us consider the problem defined by: $\Theta = [0,1]^d$, $f_t(\theta) = (1 - ||\theta - \theta_t||/\sqrt{d})^3$ where $\theta_t = t/T\,(1,\ldots,1)'$. The optimal θ^* (i.e. $\arg\max_\theta F_T(\theta)$) is $1/2\,(1,\ldots,1)'$. Figure 3.2 plots the expected regret $\sup_{\theta\in\Theta} \mathbb{E}R_T(\theta)$ (with $T = 100$, averaged over 10 experiments) as a function of the parameter N (number of sampling points/particles) for two sampling methods: the random grid mentioned in the end of Section 1.2 and the PMC method. We considered two values of the space dimension: $d = 2$ and $d = 20$. Note that the uniform discretization technique is not applicable in the case of dimension $d = 20$ (because of the curse of dimensionality). We used $K = 5$ steps and used a Gaussian centered kernel $g_{t,k}$ of variance $\sigma^2 = 0.1$ for the PMC method.

Since the complexity of sampling from a PMC method with N particles and from a grid of N points is not the same, in order to compare the performance of the two methods both in terms of regret and runtime, we plot in Figure 3.3 the regret as a function of the CPU time required to do the sampling, for different values of N.

As expected, the PMC method is more efficient since its allocation of points (particles) depends on the cumulative rewards F_t (it thus may be considered as an adaptive algorithm).

2 Applications to learning problems

2.1 Online regression

Consider an online adversarial regression problem defined as follows: at each round t, an opponent selects a couple (x_t, y_t) where $x_t \in \mathcal{X}$ and $y_t \in \mathcal{Y} \subset \mathbb{R}$, and shows the input x_t to the learner. The learner selects a regression function $g_t \in \mathcal{G}$ and predicts $\widehat{y}_t = g_t(x_t)$. Then the output y_t is revealed and the learner incurs the reward (or equivalently a loss) $l(\widehat{y}_t, y_t) \in [0, 1]$.

Since the true output is revealed, it is possible to evaluate the reward of any $g \in \mathcal{G}$, which corresponds to the full information case.

Now, consider a parametric space $\mathcal{G} = \{g_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ of regression functions, and assume that the mapping $\theta \mapsto l(g_\theta(x), y)$ is Lipschitz w.r.t. θ with a uniform (over $x \in \mathcal{X}, y \in \mathcal{Y}$) Lipschitz constant $\lambda < \infty$. This happens for example when $\mathcal{X}$ and $\mathcal{Y}$ are compact domains, the regression $\theta \mapsto g_\theta$ is Lipschitz, and the loss function $(u, v) \mapsto l(u, v)$ is also Lipschitz w.r.t. its first variable (such as for e.g. L_1 or L_2 loss functions) on compact domains.

The online learning problem consists in selecting at each round t a parameter $\theta_t \in \Theta$ such as to optimize the accuracy of the prediction of y_t with $g_{\theta_t}(x_t)$. If we define $f_t(\theta) \stackrel{\text{def}}{=} l(g_\theta(x), y)$, then applying the ALF strategy described previously (changing rewards into losses by using the transformation $u \mapsto 1 - u$), we obtain directly that the expected cumulative loss of the ALF strategy is almost as small as that of the best regression function in $\mathcal{G}$, in the sense that:

$$\mathbb{E}\Big[\sum_{t=1}^{T} l_t\Big] - \inf_{g \in \mathcal{G}} \mathbb{E}\Big[\sum_{t=1}^{T} l(g(x_t), y_t)\Big] \leqslant 2\sqrt{dT \ln(d^\alpha \lambda T)},$$

where $l_t \stackrel{\text{def}}{=} l(g_{\theta_t}(x_t), y_t)$. To illustrate, consider a feedforward neural network (NN) Bishop (2006) with parameter space Θ (the set of weights of the network) and one hidden layer. Let n and m be the number of input (respectively hidden) neurons. Thus if $x \in \mathcal{X} \subset \mathbb{R}^n$ is the input of the NN, a possible NN architecture would produce the output: $g_\theta(x) \stackrel{\text{def}}{=} \theta^o \cdot \sigma(x)$ with $\sigma(x) \in \mathbb{R}^m$ and $\sigma(x)_l \stackrel{\text{def}}{=} \sigma(\theta^i_l \cdot x)$ (where σ is the sigmoid function) is the output of the l-th hidden neuron. Here $\theta = (\theta^i, \theta^o) \in \Theta \subset \mathbb{R}^d$ the set of (input, output) weights (thus here $d = n \times m + m$).

The Lipschitz constant of the mapping $\theta \mapsto g_\theta(x)$ is upper bounded by the quantity $\sup_{x \in \mathcal{X}, \theta \in \Theta} ||x||_\infty ||\theta||_\infty$, thus assuming that the domains $\mathcal{X}$, $\mathcal{Y}$, and Θ are compacts, the assumption that $\theta \mapsto l(g_\theta(x), y)$ is uniformly (over $\mathcal{X}, \mathcal{Y}$) Lipschitz w.r.t. θ holds e.g. for L_1 or L_2 loss functions, and the previous result applies.

Now, as discussed above about the practical aspects of the ALF strategy, in this online regression problem, the knowledge of the past input-output pairs $(x_s, y_s)_{s<t}$ enables to com-

pute the weight $w_t(\theta) = \exp(\eta \sum_{s=1}^{t-1} l(g_\theta(x_s), y_s))$ for any $\theta \in \Theta$, and thus enables to use a PMC algorithm to sample θ_t from the distribution p_t. Up to our knowledge, we believe this is a first regret bound analysis of online learning for non-linear NN regression, in an adversarial setting.

2.2 Online classification

Now consider the problem of online classification (i.e. when the set of labels $\mathcal{Y}$ is finite). Here we can no longer make the assumption that the classifier's prediction $g_\theta(x) \in \mathcal{Y}$ is Lipschitz w.r.t. the parameter θ (and neither that the loss function $l(y, y') = \mathbb{I}_{\{y=y'\}}$ is Lipschitz w.r.t. its first variable). One way to circumvent this problem is to consider a class $\mathcal{G} = \{g_\theta, \theta \in \Theta\}$ of stochastic classifiers, so that $g_\theta(y|x)$ represents the probability of predicting label y given input x. The ALF strategy would apply as follows: at round t, the algorithms chooses $\theta_t \in \Theta$ and samples the prediction $\widehat{y}_t$ from the distribution $g_{\theta_t}(\cdot|x_t)$.

When the label y_t is revealed, the loss function $f_t(\theta) \stackrel{\text{def}}{=} g_\theta(y_t|x_t)$ for all classifiers g_θ may be computed. Thus assuming that the mapping $\theta \mapsto g_\theta(y|x)$ is Lipschitz w.r.t. θ with uniform (over $\mathcal{X} \times \mathcal{Y}$) Lipschitz constant λ, then Theorem 3.1 applies, and we have that

$$\underbrace{\sup_{g \in \mathcal{G}} \mathbb{E}\left\{\sum_{t=1}^{T} g(y_t|x_t)\right\}}_{\substack{\text{Exp. nb. of correct} \\ \text{predictions of best classifier}}} - \underbrace{\mathbb{E}\left\{\sum_{t=1}^{T} g_{\theta_t}(y_t|x_t)\right\}}_{\substack{\text{Exp. nb. of correct} \\ \text{predictions of ALF algo.}}} \leqslant 2\sqrt{dT \ln(cd^\alpha \lambda T)}$$

which says that the expected number of good predictions of the ALF strategy is almost as good as that of the best classifier in $\mathcal{G}$. An example of such parametric regression setting is the case of neural networks (parameterized by θ) where the activation of the output neurons (one for each label y of $\mathcal{Y}$), up to some renormalization, define the probability distribution $g_\theta(y|x)$.

2.3 Online classification with bandit information

In the previous section, the information revealed by the opponent enables to compute the reward (or loss) function $f_t(\theta)$ for all $\theta \in \Theta$. In the bandit information case considered now only the reward $f_t(\theta_t)$ of the selected action is revealed. Under our Lipschitz assumption on the functions, the knowledge of f_t at a point θ_t reveals very few information about f_t elsewhere. Thus we cannot expect to derive tight regret bounds in general. However we can obtain interesting bounds in the case when the reward function f_t may actually be coded by a finite amount of information. We illustrate this setting on the online classification problem described in Section 2.2 but with the difference that the true label $y_t \in \mathcal{Y} = \{1, \ldots, K\}$ is not revealed at each round: the only available information is $Z_t \stackrel{\text{def}}{=} \mathbb{I}_{\{\widehat{y}_t = y_t\}}$, i.e. whether the prediction $\widehat{y}_t$ is correct or not. An example of applications is the problem of web advertisement systems, where the user's click is the only received feedback.

Initialization: Set $w_1(\theta) = 1$ for all $\theta \in \Theta$.

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

(1) The adversary chooses $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and shows x_t to the learner,

(2) The learner chooses $\theta_t \sim p_t$, where $p_t(\theta) \stackrel{\text{def}}{=} \dfrac{w_t(\theta)}{\int_\Theta w_t(\theta) d\theta}$, and predicts $\widehat{y}_t \sim q_{t,\theta_t}$, where $q_{t,\theta}(y) \stackrel{\text{def}}{=} (1-\gamma) g_{\theta_t}(y|x_t) + \frac{\gamma}{K}$,

(3) The learner sees the (bandit) information $Z_t \stackrel{\text{def}}{=} \mathbb{I}_{\widehat{y}_t = y_t}$, from which he defines $\tilde{f}_t(\theta) \stackrel{\text{def}}{=} \frac{g_\theta(\widehat{y}_t|x_t)}{q_t(\widehat{y}_t)} Z_t$, where $q_t(y) \stackrel{\text{def}}{=} \int_\Theta p_t(\theta) q_{t,\theta}(y) d\theta$, for any $y \in \mathcal{Y}$.

(4) The weight function w_t is updated according to $w_{t+1}(\theta) = w_t(\theta) e^{\eta \tilde{f}_t(\theta)}$, for all $\theta \in \Theta$.

Figure 3.4: The Adversarial Lipschitz Bandit Classifier (ALBC algo)

Again, we consider a parametric family of stochastic classifiers $\mathcal{G} = \{g_\theta, \theta \in \Theta\}$, where $g_\theta(y|x)$ corresponds to the probability of selecting $y \in \mathcal{Y}$ given the input x. Now, in each round, a classifier g_{θ_t} is selected (by sampling $\theta_t \sim p_t$) and a prediction $\widehat{y}_t$ is made. However, in this bandit setting, the feedback information $Z_t = \mathbb{I}_{\{\widehat{y}_t = y_t\}}$ does not enable to evaluate the performance $f_t(\theta) \stackrel{\text{def}}{=} g_\theta(y_t|x_t)$ of any classifiers g_θ, $\theta \in \Theta$. Instead, we randomize the prediction by considering a mixture distribution between g_{θ_t} and the uniform distribution: $\widehat{y}_t \sim q_{t,\theta_t}$, where $q_{t,\theta}$ is the distribution over the labels $\mathcal{Y}$ defined by $q_{t,\theta}(y) \stackrel{\text{def}}{=} (1-\gamma) g_\theta(y|x_t) + \frac{\gamma}{K}$.

This idea is close to the Exp4 algorithm in Auer et al. (2003). Given the information Z_t, we build an estimate $\tilde{f}_t(\theta)$ of the performance $f_t(\theta)$ of any classifiers g_θ: $\tilde{f}_t(\theta) \stackrel{\text{def}}{=} \frac{g_\theta(\widehat{y}_t|x_t)}{q_t(\widehat{y}_t)} Z_t$, where $q_t(y) \stackrel{\text{def}}{=} \mathbb{E}_{\theta \sim p_t}[q_{t,\theta}(y)]$, for any $y \in \mathcal{Y}$. This estimate is unbiased since:

$$\begin{aligned}
\mathbb{E}_{\theta_t, \widehat{y}_t} \tilde{f}_t(\theta) &= \int_\Theta p_t(\theta') \sum_{y \in \mathcal{Y}} \frac{q_{t,\theta'}(y) g_\theta(y|x_t)}{q_t(y)} \mathbb{I}_{\{y = y_t\}} d\theta' \\
&= \int_\Theta p_t(\theta') \frac{q_{t,\theta'}(y) g_\theta(y_t|x_t)}{q_t(y_t)} d\theta' = g_\theta(y_t|x_t) = f_t(\theta)
\end{aligned}$$

Figure 3.4 describes this Adversarial Lipschitz Bandit Classifier (ALBC) algorithm. The next result assesses the expected performance of the ALBC algorithm $\sum_{t=1}^T \mathbb{I}_{\{\widehat{y}_t = y_t\}}$ in comparison with the expected performance of the best classifier $g \in \mathcal{G}$, in terms of number of

correct predictions. Define the regret:

$$R_T(\theta) \stackrel{\text{def}}{=} \sum_{t=1}^{T} g_\theta(y_t|x_t) - \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}_{\{\widehat{y}_t=y_t\}}\Big].$$

The ALBC algorithm has a regret $\sup_{\theta\in\Theta} \mathbb{E}R_T(\theta) \leqslant 4\sqrt{KdT\ln(cd^\alpha\lambda T)}$ (the proof follows the same lines as the proof of ALF strategy combined with EXP4 ideas). Notice that like in the multi-armed bandit problem, in this bandit setting, the regret suffers from an additional factor K per round (i.e. $\sqrt{T}$ is replaced by $\sqrt{KT}$ in the bound), compared to the full information case.

A practical algorithm. A practical implementation of the ALBC algorithm requires being able to sample θ_t from p_t. The key difference with the technique detailed in Section 1.3 is that in the ALBC algorithm, the functions $\tilde{f}_t(\theta)$ depend on $q_t(\widehat{y}_t)$ which is not directly known. However a refined MCMC or PMC algorithm is possible: at round t, assume that we have kept in memory the information: $H_{<t} \stackrel{\text{def}}{=} \{(x_s, \widehat{y}_s, Z_s, q_s(\widehat{y}_s))_{s<t}\}$.

We now show that (1) this is possible, and (2) this is sufficient for sampling $\theta_t \sim p_t$. We prove (1) recursively by showing that from $H_{<t}$ we are able to calculate $q_t(\widehat{y}_t)$ (the other pieces of information $x_t, \widehat{y}_t$, and Z_t are available at the end of round t). Thus we only need to prove that from $H_{<t}$, we can sample $\theta_t \sim p_t$ and compute $q_t(\widehat{y}_t)$. But since $q_t(\widehat{y}_t)$ is the expectation of $q_{t,\theta}(\widehat{y}_t)$ for $\theta \sim p_t$, we may consider a MCMC or PMC method where the same Markov chain (having p_t as stationnary distribution) or particle population serves for both sampling $\theta_t \sim p_t$ and estimating $q_t(\widehat{y}_t)$. Finally, this is possible since the pointwise evaluation of w_t (thus of p_t up to a renormalization constant) only depends on information in $H_{<t}$.

3 Conclusion

We have considered the adversarial online learning framework in the case of Lipschitz functions. In the full information case, the bound shows the same rate $\sqrt{dT}$ as for linear functions. This enables to derive similar performance bounds for online regression and classification, thus extending previous results to non-linear parametric approximation, such as neural networks. Our main contribution was to consider a continuous extension of the EWF algorithm (ALF strategy) for which we provide geometrical conditions for sound regret analysis, and discuss the use of different approximation schemes and especially the use of a PMC sampling method compared to non adaptive sampling methods. We provided experiments showing the benefit of using a PMC sampling method for minimizing regret under computational time constraint compared to naive random grid.

We applied this result to derive bounds for (full information) regression and classification online learning problems and (bandit information) K-classes classification problems where the revealed information is the correctness of the prediction. We derived a regret bound on

the expected number of mistakes of order $\sqrt{dTK}$, and illustrate the case of a Neural Networks architecture.

4 Proof of Theorem 3.1 (ALF strategy)

We start by following the usual proof for exponentially weighted forecasting. Define $W_t \stackrel{\text{def}}{=} \int_\Theta w_t$. For any $t \in \{1, \dots, T\}$, we have:

$$\frac{W_{t+1}}{W_t} = \frac{\int_\Theta \exp(\eta F_t)}{\int_\Theta \exp(\eta F_{t-1})} = \int_\Theta p_t(\theta) \exp(\eta f_t(\theta)).$$

Since $\exp(u) \leqslant 1 + u + u^2$ for $u \leqslant 1$, then, whenever $\eta \leqslant 1$, we have $\frac{W_{t+1}}{W_t} \leqslant 1 + \eta \int_\Theta p_t f_t + \eta^2 \int_\Theta p_t f_t^2$. Moreover, since $W_1 = \mu(\Theta)$, we get:

$$\ln(W_{T+1}) \leqslant \eta \sum_{t=1}^{T} \int_\Theta p_t f_t + T\eta^2 + \ln(\mu(\Theta)). \tag{3.4}$$

Let us write $h(\theta) \stackrel{\text{def}}{=} \exp(\eta F_T(\theta))$, and $h^* \stackrel{\text{def}}{=} \max_{x \in \Theta} h(\theta)$. We have that

$$\begin{aligned} |h(\theta_1) - h(\theta_2)| &\leqslant \eta |F_T(\theta_1) - F_T(\theta_2)| h^* \\ &\leqslant \eta \lambda T h^* ||\theta_1 - \theta_2||, \end{aligned} \tag{3.5}$$

since the function F_T is λT-Lipschitz. Let θ^* be any point of maximum of h, and define $\pi(\theta) \stackrel{\text{def}}{=} \max(0, 1 - \eta\lambda T ||\theta - \theta^*||)$. Then for all $\theta \in \Theta$,

$$h(\theta) \geqslant h^* \pi(\theta). \tag{3.6}$$

Indeed, this holds for any $\theta \notin B(\theta^*, 1/(\eta\lambda T))$ where $B(\theta, r)$ is the ball $\{x', ||x - x'|| \leqslant r\}$, since in that case, $\pi(\theta) = 0$. Now if there were some $\theta \in B(\theta^*, 1/(\eta\lambda T))$ such that $h(\theta) < h^*\pi(\theta)$, then we would have: $h(\theta^*) - h(\theta) > \eta\lambda T h^* ||x - x^*||$, which would contradict the Lipschitz property (3.5) of h.

Notice that π is a pyramid function with base $B(\theta^*, 1/(\eta\lambda T))$ and height 1. We now state a Lemma that will enable us to derive a lower bound on $\int_\Theta \pi$.

Lemma 3.1 *For any $\theta^* \in \Theta$, $r > 0$, let π be the function defined by $\pi(\theta) \stackrel{\text{def}}{=} \max(0, 1 - ||x - x^*||/r)$. Then:*

$$\int_\Theta \pi \geqslant \frac{1}{(d+1)\kappa(d)} \min\left[\mu\big(B(\theta^*, r)\big), \mu(\Theta)\right]$$

Proof:

$$\begin{aligned}\int_\Theta \pi &= \int_{\mathbb{R}^D} \mathbb{I}_{\theta\in\Theta\cap B(\theta^*,r)}(1-\frac{||\theta^*-\theta||}{r})\mu(d\theta)\\ &= \int_{\mathbb{R}^D} \mathbb{I}_{\theta\in\Theta\cap B(\theta^*,r)}\int_0^1 \mathbb{I}_{||\theta^*-\theta||\leqslant\alpha r}d\alpha\mu(d\theta)\\ &= \int_0^1\int_{\mathbb{R}^D} \mathbb{I}_{\theta\in\Theta\cap B(\theta^*,\alpha r)}\mu(d\theta)d\alpha\\ &= \int_0^1 \mu(\Theta\cap B(\theta^*,\alpha r))d\alpha\end{aligned}$$

Now, using the definition of $\kappa(d)$ from (3.1),

$$\int_\Theta \pi \geqslant \int_0^1 \frac{1}{\kappa(d)}\min[\alpha^d\mu(B(\theta^*,r)),\mu(\Theta)]d\alpha$$

We deduce that if $\mu(\Theta)\geqslant\mu(B(\theta^*,r))$ then $\int_\Theta\pi\geqslant\frac{\mu(B(\theta^*,r))}{(d+1)\kappa(d)}$. And otherwise, $\exists\alpha_0<1$ such that $\mu(\Theta)=\alpha_0^d\mu(B(\theta^*,r))$ thus we have $\int_\Theta\pi\geqslant\frac{\mu(\Theta)}{\kappa(d)}(1-\alpha_0+\frac{\alpha_0}{d+1})\geqslant\frac{\mu(\Theta)}{(d+1)\kappa(d)}$ and the Lemma is proved. □

We apply this Lemma with the π function and $r=1/\eta\lambda T$ to obtain:

$$\int_\Theta \pi \geqslant \frac{1}{(d+1)\kappa(d)}\min\big[\mu\big(B(\theta^*,\frac{1}{\eta\lambda T})\big),\mu(\Theta)\big]$$

Now using (3.6) together with the previous bound combined with Assumption A1 (i.e. $\kappa(d)\leqslant\kappa^d$ and $\mu\big(B(\theta^*,r)\big)\geqslant(r/(\kappa' d^\alpha)^d)$, we derive the lower bound:

$$\int_\Theta h \geqslant h^*\min\big[\frac{1}{(cd^\alpha\eta\lambda T)^d},\frac{\mu(\Theta)}{c^d}\big].$$

where we set $c=2\kappa\max(\kappa',1)$.

From its definition, $W_{T+1}=\int_\Theta h$, thus

$$\ln(W_{T+1})\geqslant\eta\max_{\theta\in\Theta}F_T(\theta)-\ln\left(\max\big[(cd^\alpha\eta\lambda T)^d,\frac{c^d}{\mu(\Theta)}\big]\right),$$

which, together with (3.4) yields:

$$\sup_{\theta\in\Theta}F_T(\theta)-\sum_{t=1}^T\int_\Theta p_tf_t\leqslant T\eta+\frac{1}{\eta}\max\big[d\ln(cd^\alpha\eta\lambda T)+\ln(\mu(\Theta)),d\ln c\big].$$

Since $\int_\Theta p_tf_t=\mathbb{E}_t[f_t(\theta_t)]$, where $\mathbb{E}_t$ denotes the expectation w.r.t. the choice of $\theta_t\sim p_t$, we deduce that the expected regret (w.r.t. the internal randomization of the learner) of any $\theta\in\Theta$ is bounded according to:

$$\mathbb{E}R_T(\theta)\leqslant T\eta+\frac{1}{\eta}(d\ln(cd^\alpha\eta\lambda T)+\ln(\mu(\Theta))),$$

whenever $d\ln(d^\alpha\eta\lambda T) \geqslant -\ln(\mu(\Theta))$.

Now, for the high probability result, if we introduce $Y_t = \int_\Theta p_t f_t - f_t(\theta_t)$ and $\mathcal{F}_{<t}$ the σ-algebra generated by the past random decisions, then $\mathbb{E}[Y_t|\mathcal{F}_{<t}] = 0$, thus $Y_1, ..., Y_T$ is a martingale difference sequence, and since $f_t \in [0,1]$, $|Y_t| \leqslant 1$ a.s., using Hoeffding-Azuma's inequality (see e.g. Devroye et al. (1996)), we obtain that with probability at least $1-\beta$,

$$\sum_{t=1}^{T} \int_\Theta p_t f_t \leqslant F_T + \sqrt{2T\ln(\beta^{-1})},$$

which enables to deduce (3.3).

Chapter 4

Adaptive Bandits: Towards the Best History-dependent Strategy.

In this chapter, we now consider the setting of multi-armed bandit with an adversarial environment and partial information. We consider that we have a multi-armed bandit game, thus that the environment is a possibly adaptive (but not necessarily the meanest) opponent, with the goal to design efficient algorithms for this setting. The reason not to consider only the meanest opponent is because in practice, an algorithm will not necessarily face such a bad opponent, but may also face weaker opponents. Now an algorithm designed only for the worst case may not achieve optimal performance in this setting, and it is a challenging question to design algorithms that are adaptive to the weakness of the opponent.

More precisely, we introduce models Θ of constraints based on equivalence classes on the common history (information shared by the player and the opponent) which define two learning scenarios: (1) The opponent is constrained, i.e. he provides rewards that are stochastic functions of equivalence classes defined by some model $\theta^* \in \Theta$. The regret is measured with respect to (w.r.t.) the best history-dependent strategy. (2) The opponent is arbitrary and we measure the regret w.r.t. the best strategy among all mappings from classes to actions (i.e. the best history-class-based strategy) for the best model in Θ. This allows to model opponents (case 1) or strategies (case 2) which handles finite memory, periodicity, standard stochastic bandits and other situations.

When $\Theta = \{\theta\}$, i.e. only one model is considered, we derive *tractable* algorithms achieving a *tight* regret (at time T) bounded by $\tilde{O}(\sqrt{TAC})$, where C is the number of classes of θ. Now, when many models are available, all known algorithms achieving a nice regret $O(\sqrt{T})$ are unfortunately *not tractable* and scale poorly with the number of models $|\Theta|$. Our contribution here is to provide *tractable* algorithms with regret bounded by $(TA)^{2/3}C^{1/3}\log(|\Theta|)^{1/2}$.

This work has been published in the proceedings of the *14th international conference on Artificial Intelligence and Statistics (AI&Stats 2011)*, see Maillard and Munos (2011). It also has to be related to chapter 11 where a similar notion of models is introduced but in the more general setting of reinforcement learning.

Contents

1 Introduction

Designing medical treatments for patients infected by the Human Immunodeficiency Virus (HIV) is challenging due to the ability of the HIV to mutate into new viral strains that become, with time, resistant to a specific drug Ernst et al. (2006). Thus we need to alternate between drugs. The standard formalism of *stochastic bandits* (see Robbins (1952)) used for designing medical treatment strategies models each possible drug as an arm (action) and the immediate efficiency of the drug as a reward. In this setting, the rewards are assumed to be i.i.d., thus the optimal strategy is constant in time. However in the case of adapting viruses, like the HIV, no constant strategy (i.e., a strategy that constantly uses the same drug) is good on the long term. We thus need to design new algorithms (together with new performance criteria) to handle a larger class of strategies that may depend on the whole treatment history (i.e., past actions and rewards).

More formally, we consider a sequential decision making problem with rewards provided by a possibly adaptive opponent. The general game is defined as follows: At each time-step t, the decision-maker (or player, or agent) selects an action $a_t \in \mathcal{A}$ (where $\mathcal{A}$ is a set of $A = |\mathcal{A}|$ possible actions), and simultaneously the opponent (or adversary or environment) chooses a reward function $r_t : \mathcal{A} \mapsto [0, 1]$. The agent receives the reward $r_t(a_t)$. In this paper we consider the so-called *bandit information* setting where the agent only sees the rewards of the chosen action, and not the other rewards provided by the opponent. The goal of the agent is to maximize the cumulative sum of the rewards received, i.e. choose a sequence of actions $(a_t)_{t\leqslant T}$ that maximizes $\sum_{t=1}^{T} r_t(a_t)$.

Motivating Example In order to better understand our goal, consider the following very simple problem for which no standard *bandit* algorithm achieves good cumulative rewards.

The set of actions is $\mathcal{A} = \{a, b\}$, and the opponent is defined by: $r(a) = 1$ and $r(b) = 0$ if the last action of the player is b, and $r(a) = 0$ and $r(b) = 1$ if the last action is a. Finally $r(a) = r(b) = 1$ for the first action.

In that game, playing a constant action a (or b) yields a cumulative reward of $T/2$ at time T. On the other hand, a player that would switch its actions at each round would obtain a total rewards of T, which is optimal. Although this opponent is very simple, the loss of using any usual multi-armed bandit algorithm (such as UCB Auer et al. (2002) and EXP3 Auer et al. (2003)) instead of this simple switching strategy is linear in T.

Adaptive Opponents In this paper, we consider the setting when the opponent is *adaptive*, in the sense that the reward functions can be arbitrary measurable functions of the past common history, where by common history we mean all the observed rewards $(r_s(a_s))_{s<t}$ and actions $(a_s)_{s<t}$ played before current time t. We write $h_{<t}$ or simply h the common history up to time t, so we can write $r_t(a) = r(h_{<t}, a)$.

Due to the motivating example, we naturally want to compare to the best history-dependent strategy against the adaptive opponent, and introduce a more challenging notion of regret (see Section 2.1) than usual for that purpose. Since this may be not possible without assumptions on the opponent or the comparison strategies (see Ryabko and Hutter (2008)), we consider some model of constraints, and thus we want to adapt to a class Θ of possible constraints. The question is: can we adapt to the (unknown) model of constraints of the opponent?

Adversarial Bandits In Literature A first approach when considering adversarial bandits providing arbitrary rewards (when no constraint is put on the complexity of the adversary) is to assess the performance of the player in terms of the best strategy that is constant in time (best constant action), which defines the external regret Auer et al. (2003), Freund and Schapire (1995). However, since this approach does not consider limitations on the strategy of the opponent w.r.t. the history, it can only give partial answer to the question of adaptivity to the best possible history-dependent strategy against a given opponent.

In Auer et al. (2003), the authors extend the class of comparison strategies to piecewise constant strategies with at most S switches. The corresponding Exp3S (aka ShiftBand) algorithm achieves a regret of order $\sqrt{TSA\log(T^3A)}$, provided that T is large enough. However, against the opponent described in the previous section, the best strategy would need to switch $S = T/2$ times, thus this algorithm still suffers a linear regret compared to the optimal strategy.

The notion of internal regret (see Foster and Vohra (1996)), which compares the loss of an online algorithm to the loss of a modified algorithm that consistently replaces one action by another, has been also considered in many works Hart and Mas-Colell (2000), Stoltz (2005), Cesa-Bianchi and Lugosi (2003), Foster and Vohra (1999). Following the work of Lehrer and Rosenberg (2003), in Blum and Mansour (2005) the authors propose a way to convert any external regret minimization algorithm into an algorithm minimizing an extended notion of internal regret, using the so-called modifications rules that are functions $h, a \to b$, where h is the history, and a and b are actions, see also Blum and Mansour (2007). This enables to compare the actions selected by the algorithm to an alternative sequence and thus to assess

the performance of the algorithm to other slightly perturbed algorithm. Assuming that the opponent's strategy can be described with the modification rules, then we might also see the corresponding modified regret minimization algorithm as adaptive to the opponent, in some sense. However, the proposed algorithm uses exponentially many internal variables and will not provide tight performance bounds in terms of regret w.r.t. the best history-based strategy, that we consider in Section 2.1.

On a more theoretical aspect, the work by Ryabko and Hutter (2008) addresses the learnability problem in reactive environments (adaptive opponents). The authors introduce the notion of value-stable and recoverable environments, and show that environments satisfying such mild conditions are learnable. This also means that it is *not possible* to obtain sublinear regret w.r.t. the best strategy against any arbitrary opponent: we need to consider limitations of the opponent. Note also that the main proof of the paper by Ryabko and Hutter (2008) is constructive, but unfortunately the would-be corresponding player is not implementable.

Tractability Since bandit algorithms are the base stone for Reinforcement Learning (RL) algorithms, it is thus important if not crucial to consider numerically efficient algorithms. The question of adaptability is not trivial because of tractability: Although the works of Blum and Mansour (2005) and Ryabko and Hutter (2008) already provide adaptive algorithms, none of them would be tractable in our setting (even with only one θ). Moreover, for a pool of possible behaviors Θ of the opponent (see Section 3), we define the Θ-regret w.r.t. the best possible strategy for the best model $\theta \in \Theta$. We then show (in Section 3) that our problem can be seen as a special instance of sleeping bandits. The best regret bounds known with tractable algorithms would be of order $\tilde{O}((TC_\Theta)^{4/5})$ (see Kanade et al. (2009)) whereas there exists a non-tractable algorithm achieving $\tilde{O}(\sqrt{TC_\Theta})$, where $C_\Theta = \sum_{\theta\in\Theta} C_\theta$ and C_θ is the complexity of model θ. If the regret of the second algorithm nicely scales with the time horizon T, both of them provide loose bounds for large $|\Theta|$. So the question is: can we design *tractable* algorithms that can *adapt* to a *large* pool of models of constraints?

Our Contribution The main contribution of this paper is a new way of considering adversarial opponents. For some equivalence relation Φ on histories, we write $[h]_\Phi$ for the equivalence class of the history h w.r.t. Φ. We introduce Φ-constrained opponents that are such that the reward functions only depend on the equivalence classes of history, i.e. $r_t(a) = r([h_{<t}], a)$. Similarly, one can consider classes of strategies of the form $\mathcal{H}/\Phi \mapsto \mathcal{A}$, where $\mathcal{H}/\Phi$ is the set of equivalence classes of histories. Interestingly, such equivalence relations were also introduced in Hutter (2009), with the goal to build relevant equivalence relations for Reinforcement Learning. The author provides useful insights, but no performance analysis. Our model of constraints, although seemingly simple, has two main advantages: (1) the notion of Φ-regret (see Section 2) captures the regret w.r.t. such strategies and is expressive enough to handle many kinds of situations (like finite memory, periodicity, etc) and thus enables to define opponents that can be anything *from the worst possible* (fully adversarial), *to a simple stochastic* multi-armed bandit. (2) such a model leads to simple

and efficient algorithms that are built directly from standard algorithms, and yet achieve significantly good performances.

The introductory Section 2 starts with a single model and provides algorithms with expected regret w.r.t. the optimal history-based strategy bounded by $O(\sqrt{TAC\log A})$, where C is a measure of the complexity (number of equivalence classes of $\mathcal{H}/\Phi$) of the opponent, and a lower bound $\Omega(\sqrt{TAC})$. This applies to the switching opponent described in the introduction. The complexity of those algorithms is C times the complexity of the standard algorithms they are built from (namely UCB and Exp3), as opposed to the complexity of order A^C for algorithms that would be derived directly from Blum and Mansour (2005) in our setting. Note also that for the special case of a Φ-constrained opponent with a known model Φ, one can consider a RL point of view instead, and apply algorithms such as Ortner (2009).

Our main contribution in this paper is to consider the more challenging situation where we have a pool of possible models Θ. In this case, we provide tractable algorithms with Θ-regret of order (see Section 3). $(TA)^{2/3}(C_{\theta^*}\log(T))^{1/3}\log(|\Theta|)^{1/2}$ when the opponent belongs to the pool (i.e. $\theta^*\in\Theta$, in which case we compare the performance to that of the optimal history-based strategy), and $T^{2/3}(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}$ where $\overline{C}=\max_\theta C_\theta$, when it does not (in which case we compare to the best $\mathcal{H}/\Phi_\theta$-history-class-based strategy for the best model $\theta\in\Theta$).

We finally report numerical experiments in Section 4 which compares standards algorithms for bandits (from stochastic to adversarial) (UCB, MOSS, EXP3, ShiftBand) to the algorithms proposed here.

2 Preliminary results

2.1 Model of constrained opponents

Let $\mathcal{H}$ be the set of all histories, i.e. sequences of action played and information received. Let $\Phi:\mathcal{H}\to Y$ be a given function mapping histories to an abstract space Y, and let $\mathcal{H}/\Phi$ denote the class of equivalence of histories w.r.t. the relation $h_1\sim h_2$ if and only if $\Phi(h_1)=\Phi(h_2)$. We write also $[h]_\Phi$ (or $[h]$ when there is no ambiguity) for the equivalence class of h.

Based on an equivalence-class Φ, one can define Φ-constrained opponents, which are intuitively the opponents that are Φ-classwise stochastic:

Definition 4.1 (Φ-*constrained opponent*) *A Φ-constrained opponent is a function $f:\mathcal{H}/\Phi\to\Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the set of distribution over the set $\mathcal{A}$, taking values in $[0,1]$ (i.e. we assume that all rewards belongs to the interval $[0,1]$).*

Examples: Definition 4.1 covers many situations:

- When $\Phi(h)=1$ for all $h\in\mathcal{H}$, then $\mathcal{H}/\Phi$ consists of only one class, and Definition 4.1 reduces to a stochastic multi-armed bandit.

- When $\Phi_m : \mathcal{H} \to \mathcal{A}^m$ is $\Phi_m(h) = a_1...a_m$, where $a_1, ..., a_m$ are the last sequence of m actions, this corresponds to opponents with finite short-term memory of length m. In this case, there are $|\mathcal{A}|^m$ equivalence classes. The example of the introduction corresponds to this case with $m = 1$.
- When $\Phi : \mathcal{H} \to \{0, ..., m-1\}$ is defined by $\Phi(h) = |h| \bmod m$, where $|h|$ is the length of the history in term of number of time steps, it corresponds to reward functions that come from time-periodic distributions. Here, there are m different classes.

Regret Against The Best History-class-based Strategy: If we consider a Φ**-constrained** opponent, then one can define for each class $c \in \mathcal{H}/\Phi$, and action $a \in \mathcal{A}$ the expected reward $\mu_c(a) = \mathbb{E}[r(c, a)]$. We define the expected history-class-based regret for the equivalence class defined by Φ, also called *stochastic* Φ*-regret*, as:

$$R_T^\Phi = \mathbb{E}\Big(\sum_{t=1}^{T} \max_{a \in \mathcal{A}} \mu_{[h_{<t}]}(a) - \mu_{[h_{<t}]}(a_t)) \Big), \tag{4.1}$$

where $(a_t)_{t \leqslant T}$ is the sequence of actions played, $h_{<t}$ is the history observed by the player up to time t, and $\max_a \mu_{[h_{<t}]}(a)$ is the best action, which respect to the expected rewards provided by the opponent, given the history-class $[h_{<t}]$.

Now, for an **arbitrary** adversary and an equivalence class Φ, one can define a (non-stochastic) regret w.r.t. the best $\mathcal{H}/\Phi$-history-class-based strategy, also called *adversarial* Φ*-regret*,

$$\tilde{R}_T^\Phi = \sup_{g : \mathcal{H}/\Phi \to \mathcal{A}} \mathbb{E}\Big(\sum_{t=1}^{T} \Big[r_t(g([h_{<t}])) - r_t(a_t) \Big] \Big), \tag{4.2}$$

where $g([h_{<t}])$ is the action that a strategy g would play given the history-class $[h_{<t}]$ activated at time t, where g belongs to the set of strategies that are constant per $\mathcal{H}/\Phi$-history-class, i.e. mappings $\mathcal{H}/\Phi \to \mathcal{A}$.

In both cases, the expectation is w.r.t. all sources of randomness: the possible internal randomization of the player, and the possible random rewards provided by the opponent.

Note that by definition the Φ-regret is always bigger than the external regret (i.e. w.r.t. the best constant action), and that in the case when Φ defines only one class, those two notions of regret reduce to their usual definitions in stochastic and adversarial bandits, respectively.

2.2 Upper bounds on the Φ-regret

In the case we play against a constrained opponent, we observe from the definition of the Φ-regret (4.1) that if we introduce $R_T(c) = \mathbb{E}\big[\sum_{t=1}^{T} \big(\max_a \mu_{[h_{<t}]}(a) - \mu_c(a_t) \big) \mathbb{I}_{[h_{<t}]=c} \big]$ for a class $c \in \mathcal{H}/\Phi$, then $R_T^\Phi = \sum_{c \in \mathcal{H}/\Phi} R_T(c)$. This enables us to use usual stochastic bandit algorithms, such as UCB Auer et al. (2002), per history-class, and the resulting behavior will enable to minimize the stochastic Φ-regret.

Similarly, if we consider an arbitrary opponent, and an equivalence class Φ, by using usual adversarial bandit algorithms, such as Exp3 Auer et al. (2003), per history-class, one can minimize the per-class regret $\mathbb{E}\big[\sum_{t=1}^{T}\big(r_t(g(c))-r_t(a_t)\big)\mathbb{I}_{[h_{<t}]=c}\big]$ w.r.t. any constant-per-class strategy g, thus minimizing the adversarial Φ-regret $\tilde{R}_T^{\Phi}$.

The two corresponding algorithms, called respectively Φ-UCB and Φ-EXP3, are described in Figure 4.1 (α and η are parameters) and we report the regret upper-bounds in the next result.

Theorem 4.1 (Φ-*regret performance bounds*) *In the case of a Φ-constrained opponent, using the Φ-UCB algorithm with parameter $\alpha > 1/2$, we have the distribution-dependent bound:*

$$R_T^{\Phi} \leqslant \sum_{c\in\mathcal{H}/\Phi;\mathbb{E}(N_T(c))>0}\ \sum_{a\in\mathcal{A};\Delta_c(a)>0} \frac{4\alpha\log(T)}{\Delta_c(a)} + \Delta_c(a)c_\alpha$$

where $N_T(c)=\sum_{t=1}^{T}\mathbb{I}_{[h_{<t}]=c}$, the per-class gaps $\Delta_c(a)\stackrel{\text{def}}{=}\sup_{b\in\mathcal{A}}\mu_c(b)-\mu_c(a)$, and the constant $c_\alpha=1+\frac{4}{\log(\alpha+1/2)}(\frac{\alpha+1/2}{\alpha-1/2})^2$. We also have a distribution-free bound (i.e. which does not depend on the gaps):

$$R_T^{\Phi} \leqslant \sqrt{TA\overline{C}\Big(4\alpha\log(T)+c_\alpha\Big)}$$

where $\overline{C}=|\{c\in|\mathcal{H}/\Phi|;\mathbb{E}(N_T(c))>0\}|$ is the number of classes that may be activated.

Now, in the case of an arbitrary opponent, using Φ-Exp3 algorithm, we have:

$$\tilde{R}_T^{\Phi} \leqslant \frac{3}{\sqrt{2}}\sqrt{T\overline{C}A\log(A)}.$$

The proof of these statements is reported in the supplementary material and directly derives from the analysis detailed in Bubeck (2010) and the previous remarks. Note that one can use other bandit algorithms (such as UCB-V Audibert et al. (2009), MOSS Audibert and Bubeck (2009)) and derive straightforwardly the corresponding result for the Φ-regret.

2.3 Lower bounds on the Φ-regret

We now derive lower bounds on the Φ-regret to show that the previous upper bounds are tight.

Intuitively, on each class c, one may suffer a regret of order $\sqrt{N_T(c)A}$, where $N_T(c)$ is the number of times class c is visited. Now, since the way classes are "visited" depends on the structure of the game and the strategy of both the player and the opponent, those classes cannot be controlled by the player only. Thus we show that there always exist an environment such that whatever the strategy of the player is, a particular opponent will lead to visit all history-classes uniformly in expectation.

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

(1) Define $\widehat{\mu}_{t,c}(a) = \frac{1}{N_{t-1}(c,a)} F^c_{t-1}(a)$, where $F^c_t(a) = \sum_{s=1}^t r_s(a) \mathbb{I}_{[h_{<s}]=c} \mathbb{I}_{a_s=a}$ and $N_t(c,a) = \sum_{s=1}^t \mathbb{I}_{[h_{<s}]=c} \mathbb{I}_{a_s=a}$.

(2) Define $\tilde{\mu}_{t,c}(a) = \widehat{\mu}_{t,c}(a) + \sqrt{\frac{\alpha \log(N_t(c))}{N_{t-1}(c,a)}}$.

(3) Compute $c_t = \Phi(h_{<t})$.

(4) Play $a_t \in \arg\max_{a \in \mathcal{A}} \tilde{\mu}_{t,c_t}(a)$.

Initialization: Define $\forall a \in \mathcal{A} \quad \xi_1(a) = \frac{1}{A}$
For each round $t = 1, 2, \ldots, T$

(1) Play $a_t \sim \xi_t$, observe $r_t(a_t)$.

(2) Define $\tilde{l}_t(c,a) = \frac{1 - r_t(a_t)}{\xi_t(a)} \mathbb{I}_{a_t=a} \mathbb{I}_{c=[h_{<t}]}$.

(3) Define $w^c_{t+1}(a) = \exp(-\eta \sum_{s=1}^t \tilde{l}_s(c,a))$.

(4) Compute $c_{t+1} = \Phi(h_{<t+1})$.

(5) Define $\xi_{t+1}(a) = \frac{w^{c_{t+1}}_{t+1}(a)}{\sum_a w^{c_{t+1}}_{t+1}(a)}$.

Figure 4.1: Φ-UCB (top) and Φ-Exp3 (down)

We consider here, for a given class function Φ, players that may depend on Φ and opponents that may depend both on Φ and on the player. Then we consider the worst opponent for the best player over the worst class-function Φ of given complexity (expressed in terms of number of classes C of $\mathcal{H}/\Phi$). The following result easily follows from Bubeck (2010).

Theorem 4.2 (Φ-*regret lower bound*) *Let* sup *represents the supremum taken over all* Φ*-constrained opponents and* inf *the infimum over all players, then the stochastic* Φ*-regret is lower-bounded as:*

$$\sup_{\Phi;|\mathcal{H}/\Phi|=C} \inf_{algo} \sup_{\Phi-opp} R^\Phi_T \geqslant \frac{1}{20}\sqrt{TAC}.$$

Let sup *represents the supremum taken over all possible opponents, then the adversarial* Φ*-regret is lower-bounded as:*

$$\sup_{\Phi;|\mathcal{H}/\Phi|=C} \inf_{algo} \sup_{opp} \tilde{R}^\Phi_T \geqslant \frac{1}{20}\sqrt{TAC}.$$

3 Playing against an opponent using a pool of models

After this introductory section, we now turn to the main challenge of this paper. When playing against a given opponent, its model of constraints Φ may not be known. It is thus natural to consider several equivalence relations defined by a pool of class functions (models) $\Phi_\Theta = (\Phi_\theta)_{\theta\in\Theta}$, and that the opponent plays with some model induced by some Φ^*. We consider two cases: either $\Phi^* = \Phi_{\theta^*} \in \Phi_\Theta$, i.e. the opponent is a Φ_{θ^*}-constrained opponent with $\theta^* \in \Theta$, or the opponent is arbitrary, and we will compare our performance to that of the best model in Θ.

We define accordingly two notions of regret: If we consider a Φ^*-constrained opponent, where $\Phi^* \in \Phi_\Theta$, then one can define the so-called *stochastic* Φ_Θ*-regret* as:

$$R_T^\Theta = \mathbb{E}\Big(\sum_{t=1}^{T} \max_{a\in\mathcal{A}} \mu_{[h_{<t}]*}(a) - \mu_{[h_{<t}]*}(a_t)\Big). \tag{4.3}$$

where $[h_{<t}]*$ is the history-class used by the opponent.

Now, for an arbitrary opponent and a pool of equivalence classes Φ_Θ, we define a regret w.r.t. the best $\mathcal{H}/\Phi_\theta$-history-class-based strategy for the best model $\theta \in \Theta$, also called *adversarial* Φ_Θ*-regret*:

$$\tilde{R}_T^\Theta = \sup_{\theta\in\Theta} \sup_{g:\mathcal{H}/\Phi_\theta\to\mathcal{A}} \mathbb{E}\Big(\sum_{t=1}^{T} \Big[r_t(g([h_{<t}]_\theta)) - r_t(a_t)\Big]\Big), \tag{4.4}$$

where the class $[h_{<t}]_\theta$ corresponds to the model θ.

Tractability This problem can be seen as a Sleeping bandits (Kleinberg et al. (2008), Kanade et al. (2009)) with stochastic availability and adversarial rewards. Indeed, by considering each class c in each model θ, we get a total of $C_\Theta = \sum_{\theta\in\Theta} C_\theta$ experts. Now at each time step, only one class per model is awake, and thus the best awake expert changes with time. Recasting this problem in a usual bandit setting where the best expert is constant over time requires considering the $C_\Theta!$ possible rankings (see Kleinberg et al. (2008)), each ranking being now seen as an expert. Running Exp4 algorithm on top of this new experts would give a sleeping-bandit regret (and thus a Φ_Θ-regret) of order $O(\sqrt{TA\log(C_\Theta!)}) = O(\sqrt{TAC_\Theta\log(C_\Theta)})$. Unfortunately this algorithm is intractable and the bound is very loose when the number of models is large. In Kanade et al. (2009), they proposed a (tractable) algorithm that would achieve in our setting a regret bounded by $O((TC_\Theta)^{4/5}\log(T))$.

We now describe tractable algorithms with regret upper-bounded by $O((TA)^{2/3}\log(|\Theta|)^{1/2})$ for both the stochastic and adversarial Φ_Θ-regret, which improves upon previous bounds for our setting.

EXP4/UCB And EXP4/EXP3 Algorithms: A natural approach is to consider each model $\theta \in \Theta$ as one expert defined by a equivalence function Φ_θ and then run the Exp4 meta-algorithm (see Auer et al. (2003)) to select an action based on the recommendations of all experts. More precisely, at each time t, the meta algorithm plays a_t according to a distribution $q_t(\cdot) = \sum_\theta p_t(\theta)\xi_t^\theta(\cdot)$ which is a mixture of distributions ξ_t^θ that each expert θ assigns to each action, weighted by a distribution $p_t(\theta)$ over the set of experts Θ. Figure 4.2 describes the Exp4 algorithm (see Auer et al. (2003)) using a mixing parameter $\gamma > 0$.

Initialization: Define $\forall \theta \in \Theta, \quad p_1(\theta) = \frac{1}{|\Theta|}$.

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

(1) Define $q_t(a) = (1-\gamma)\sum_{\theta\in\Theta} p_t(\theta)\xi_t^\theta(a) + \frac{\gamma}{A}$.

(2) Draw $a_t \sim q_t$, and observe $r_t(a_t)$.

(3) Define $\tilde{l}_t(a) = \frac{1-r_t(a_t)}{q_t(a)}\mathbb{I}_{a_t=a}$.

(4) Define $g_t(\theta) = \sum_a \xi_t^\theta(a)\tilde{l}_t(a)$.

(5) Define $w_{t+1}(\theta) = \exp(-\gamma\sum_{s=1}^t g_t(\theta)/A)$.

(6) Define $p_{t+1}(\theta) = \frac{w_{t+1}(\theta)}{\sum_\theta w_{t+1}(\theta)}$.

Figure 4.2: The Exp4 meta algorithm

In Auer et al. (2003) the authors relate the performance of the meta algorithm to that of any individual expert (see Theorem 7.1 in Auer et al. (2003)). However, it is not obvious to build an algorithm for each individual expert Φ_θ that will minimize its Φ_θ-regret. Indeed, the actions played by the meta algorithm *differ* from the ones that would have been played by each specific expert θ. This means that for each expert, not only we have a limited (bandit) information w.r.t. the reward function, but also *each expert does not see the reward of its recommended action*. This results in individual expert algorithms with poorer regret bounds than in the single model case described in the previous section (for which one observes the reward of the chosen action).

We provide two algorithms based respectively on UCB and Exp3, that may be used by each individual expert θ:

- Φ_θ-UCB is defined as before (see Figure 4.1), except that instead of step (4) we define ξ_t^θ as a Dirac distribution at the recommended action $\arg\max_{a\in\mathcal{A}} \tilde{\mu}_{t,c_t}(a)$.
- Φ_θ-Exp3 is defined as before (see Figure 4.1), except that in step (1), no action is drawn from ξ_t (since the meta algorithm chooses $a_t \sim q_t$), and step (2) is replaced by:

$\tilde{l}_t(c,a) = \frac{1-r_t(a_t)}{q_t(a)}\mathbb{I}_{a_t=a}\mathbb{I}_{c=[h<t]_\theta}$ (i.e. we re-weight by using the probability $q_t(a)$ of the meta algorithm instead of the probability $\xi_t^\theta(a)$ of the individual expert θ).

Regret bounds (proved in Appendix 5.1.1,5.1.2) of the meta algorithm Exp4 combined respectively with individual algorithms UCB and Exp3 (called respectively Exp4/UCB and Exp4/Exp3) are given below.

Theorem 4.3 (Φ_Θ-regret performance bounds) *Assume that we consider a Φ^*-constrained opponent with $\Phi^* \in \Phi_\Theta$, then the stochastic Φ_Θ-regret of Exp4/UCB is bounded as:*

$$R_T^\Theta = O\Big((TA)^{2/3}(\overline{C}\log(T))^{1/3}\log(|\Theta|)^{1/2}\Big),$$

where $\overline{C} = |\mathcal{H}/\Phi^*|$ is the number of classes of the model Φ^* of the opponent. Now, for any opponent, the adversarial Φ_Θ-regret of Exp4/Exp3 is bounded as

$$\tilde{R}_T^\Theta = O\Big(T^{2/3}(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}\Big),$$

where $\overline{C} = \max_{\theta\in\Theta}|\mathcal{H}/\Phi_\theta|$ is the maximum number of classes for models $\theta \in \Theta$.

Note that, like in EXP4, we obtain a logarithmic dependence on $|\Theta|$ since playing an action that has been chosen from a mixture of the probability distributions (over actions) of all models yields a reward which provides information about all the models.

4 Experiments

We illustrate our approach with three different adaptive opponents and compare the results of standard algorithms to the algorithms described here using two measures of performance: the Φ-regret, and the external regret.

We consider only two actions $\mathcal{A} = \{a, b\}$, and fix the time horizon at $T = 500$. The three considered opponents have finite short-term memory of length $m = 0, 1, 2$ respectively, i.e. are Φ_m-constrained opponents in the sense of Definition 4.1. More precisely, the reward distributions are Bernoulli, and the opponents are

- O_0 is a simple stochastic bandit (no memory). We choose $\mu(a) = 0.4$ and $\mu(b) = 0.7$
- O_1 provides a mean reward 0.8 when the action changes at each step, and 0.3 otherwise,
- O_2 provides a mean reward 0.8 when the action changes every two steps and 0.3 otherwise.

Each plot of Figure 4.3 corresponds to one opponent (O_0 is left, O_1 right, and O_2 is bottom). In each plot, we represent the external regret (cyan) and Φ-regret (red) obtained for several algorithms. From left to right, the first four algorithms are UCB, MOSS, Exp3

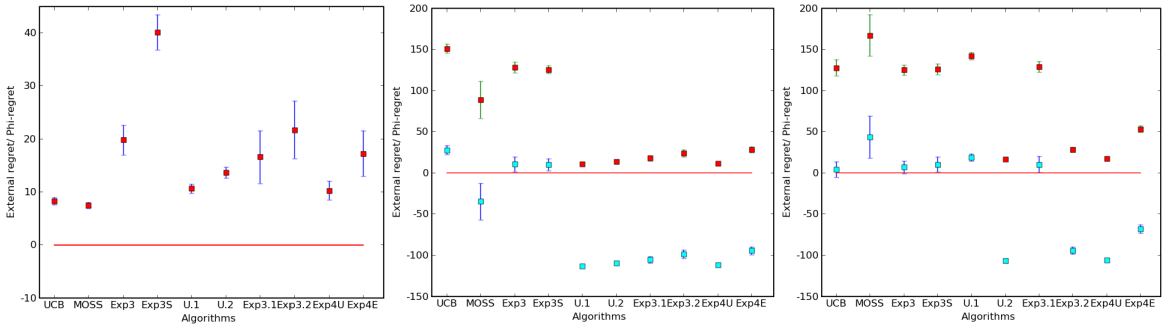


Figure 4.3: Regret w.r.t. the best history-dependent strategy (red) and best constant strategy (cyan) for 3 opponents. All experiments have been averaged over 50 trials.

and ShiftBand. The next four correspond respectively to Φ_1-UCB, Φ_2-UCB, Φ_1-Exp3, and Φ_2-Exp3 algorithms (i.e. versions of UCB and Exp3 with memory of length 1 and 2). Note that Φ_0-UCB (resp. Exp3) is just UCB (resp. Exp3). The last two algorithms correspond to the Exp4/UCB (resp. Exp4/Exp3), i.e. meta algorithm Exp4 run on top of Φ_m-UCB (resp Φ_m-Exp3) algorithms, for $m = 0, 1, 2$) as defined in Section 3.

The last two algorithms do not know the model of constraint corresponding to the opponent they are facing and still, they clearly outperform other standard algorithms (with frankly negative external regret) for the two adapting opponents (second and third). This clear improvement appears also when the model considered by the algorithm is *more* complex than that of the opponent (e.g. Φ_2-UCB facing opponent 1). On the other hand, the reverse is false (Φ_1-UCB and Φ_1-Exp facing opponent 2) since a algorithm using a piece of history of length 1 cannot play well against an opponent with memory 2.

Future work

We do not know whether in the case of a pool of models Φ_Θ, there exist tractable algorithms with Φ_Θ-regret better that $T^{2/3}$ with log dependency w.r.t. $|\Theta|$. Here we have used a meta Exp4 algorithm, but we could have used other meta algorithms using a mixture $q_t(a) = \sum_\theta p_t(\theta)\xi_t^\theta(a)$ (where the p_t are internal weights of the meta algorithm). However, when computing the approximation term of the best model θ^* by models $\theta \in \Theta$ (see the supplementary material), it seems that the Φ_Θ-regret cannot be strongly reduced without making further assumptions on the structure of the game, since in general the mixture distribution q_t may not converge to the distribution ξ_t^θ proposed by the best model $\theta \in \Theta$. This question remains open.

5 Technical details

We first remind the result of Auer et al. (2003) relating the cumulative reward of the Exp4 algorithm to the one of the best expert on top of which it is run. We have:

Lemma 4.1 *For any $\gamma \in (0,1]$, for any family of experts which includes the uniform expert, one has*

$$\max_\theta \sum_{t=1}^T \mathbb{E}_{a\sim\xi_t^\theta} r_t(a) - \mathbb{E}_{a_1,\ldots,a_T}\Big(\sum_{t=1}^T r_t(a_t)\Big) \;\leqslant\; (e-1)\gamma T + \frac{A\log(|\Theta|)}{\gamma}.$$

In our case, since the ξ_t^θ are not fixed in advance but are random variables, we can not apply the original result of Auer et al. (2003) for fixed expert advises, but need to adapt it. The proof of the following adapttion easily follows from the original proof of Theorem 7.1 in Auer et al. (2003):

Lemma 4.2 *For any $\gamma \in (0,1]$, for any family of experts which includes the uniform expert such that all expert advices are adapted to the filtration of the past, one has*

$$\max_\theta \sum_{t=1}^T \mathbb{E}_{a_1,\ldots,a_{t-1}}\Big(\mathbb{E}_{a\sim\xi_t^\theta} r_t(a)\Big) - \mathbb{E}_{a_1,\ldots,a_T}\Big(\sum_{t=1}^T r_t(a_t)\Big) \;\leqslant\; (e-1)\gamma T + \frac{A\log(|\Theta|)}{\gamma}.$$

Proof: Indeed, by construction of the algorithm, the beginning of the original proof from Auer et al. (2003) applies and gives

$$\sum_{t=1}^T r_t(a_t) \;\geqslant\; (1-\gamma)\sum_{t=1}^T \mathbb{E}_{a\sim\xi_t^\theta}\tilde{r}_t(a) - \frac{A\log(|\Theta|)}{\gamma} - (e-2)\frac{\gamma}{A}\sum_{t=1}^T\sum_{a\in\mathcal{A}}\tilde{r}_t(a),$$

where we introduce the notation $\tilde{r}_t(a) \stackrel{\text{def}}{=} \frac{r_t(a_t)}{q_t(a)}\mathbb{I}_{a_t=a}$.

Now, we use the fact that $\xi_t^\theta(a)$ is adapted to the filtration of the past $\mathcal{F}^{t-1}$ together with the property that $\mathbb{E}(\tilde{r}_t(a)|\mathcal{F}^{t-1}) = \mathbb{E}(r_t(a)|\mathcal{F}^{t-1})$ to deduce successively that

$$\begin{aligned}
\mathbb{E}\Big(\mathbb{E}_{a\sim\xi_t^\theta}\tilde{r}_t(a)\Big) &= \mathbb{E}\Big(\sum_{a\in\mathcal{A}}\mathbb{E}(\tilde{r}_t(a)\xi_t^\theta(a)|\mathcal{F}^{t-1})\Big)\\
&= \mathbb{E}\Big(\sum_{a\in\mathcal{A}}\mathbb{E}(\tilde{r}_t(a)|\mathcal{F}^{t-1})\xi_t^\theta(a)\Big)\\
&= \mathbb{E}\Big(\sum_{a\in\mathcal{A}}\mathbb{E}(r_t(a)|\mathcal{F}^{t-1})\xi_t^\theta(a)\Big)\\
&= \mathbb{E}\Big(\sum_{a\in\mathcal{A}}\mathbb{E}(r_t(a)\xi_t^\theta(a)|\mathcal{F}^{t-1})\Big)\\
&= \mathbb{E}\Big(\mathbb{E}_{a\sim\xi_t^\theta} r_t(a)\Big)
\end{aligned}$$

On the other hand, since by assumption the uniform expert belongs to the set of considered experts, we also have

$$\frac{1}{A}\mathbb{E}\Big(\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\tilde{r}_t(a)\Big) \;=\; \sum_{t=1}^{T}\mathbb{E}\big(\mathbb{E}_{a\sim U(\mathcal{A})}r_t(a)\big) \;\leqslant\; \max_{\theta}\sum_{t=1}^{T}\mathbb{E}\big(\mathbb{E}_{a\sim\xi_t^\theta}r_t(a)\big)\,,$$

where $U(\mathcal{A})$ denotes the uniform distribution over the set of actions $\mathcal{A}$. This concludes the proof. □

5.1 The rebel-bandit setting

We now introduce the setting of Rebel bandits that may have its own interest. It will be used to compute the model-based regret of the Exp4 algorithm. In this setting, we consider that at time t the player θ proposes a distribution of probability ξ_t^θ over the arms, but he actually receives the reward corresponding to an action drawn from another distribution, q_t, the distribution of probability proposed by the meta algorithm.

Following (4.4), we define the best model of the pool:

$$\theta^* = \arg\max_{\theta\in\Theta}\sup_{g:\mathcal{H}/\Phi\to\mathcal{A}}\mathbb{E}\Big(\sum_{t=1}^{T}\Big[r_t(g([h_{<t}]_\theta)) - r_t(a_t)\Big]\Big).$$

We then define for any class $c\in\mathcal{H}/\Phi_{\theta^*}$, the action $a_c^* \stackrel{\text{def}}{=} \arg\max_a \mu_c(a)$ that corresponds to the best history-class-based strategy. Thus we can also write $a_t^* = a^*_{[h_{<t}]_{\theta^*}}$. We now analyze the ($\Phi$-constrained) Exp3 and UCB algorithms in this setting and bound the corresponding rebel-regret:

Definition 4.2 (*Rebel regret*) *The Rebel-regret of the algorithm that proposes at time t the distribution ξ_t^θ but in the game where the action $a_t\sim q_t$ is played instead is:*

$$\mathcal{R}_T^q(\theta) = \sum_{t=1}^{T}\mathbb{E}_{a_1,..,a_{t-1}}\Big(r_t(a^*_{[h_{<t}]_{\theta^*}}) - \mathbb{E}_{a\sim\xi_t^\theta}(r_t(a))\Big).$$

5.1.1 Φ-Exp3 in the rebel-bandit setting

We now consider using Exp4 on top of Φ-contrained algorithms. We first use the experts Φ_θ-Exp3 for $\theta\in\Theta$ with a slight modification on the definition of the function $\tilde{l}_t(c,a)$. Indeed since the action a_t are drawn according to the meta algorithm and not Φ_θ-Exp3, we redefine $\tilde{l}_t(c,a) = \frac{1-r_t(a)}{q_t(a)}\mathbb{I}_{a_t=a}\mathbb{I}_{[h_{<t}]_\theta=c}$ so as to get unbiased estimate of $r_t(a)$ for all a. We now provide a bound on the Rebel-regret of the Φ^*-Exp3 algorithm.

Theorem 4.4 (*Rebel regret bound for Exp3*) *The Φ_{θ^*}-Exp3 algorithm in the Rebel bandit setting where $q_t(a) \geqslant \delta$ for all a, and choosing the parameter $\eta^\theta_{t^\theta_c(i)} = \sqrt{\frac{\delta \log(A)}{i}}$ satisfies*

$$\mathcal{R}^q_T(\theta^*) \leqslant 2\sqrt{\frac{T\overline{C}\log A}{\delta}}$$

Proof: The proof is in six steps and mainly follows the proof in Section 2.1 of Bubeck (2010) that provides a bound on the regret of Exp3 algorithm.

Since we only consider the model θ^*, we will simply refer to it as θ and also write c_t for $[h_{<t}]_{\theta^*}$ to avoid cumbersome notations.

Step 1. Rewrite the regret term to make appear the actions a_t chosen by the meta algorithm at time t. By definition of $\tilde{l}^\theta_t(c_\theta, a)$ we have $\mathbb{E}_{a_t\sim q_t}(\tilde{l}^\theta_t(c_t, a)) = 1 - r_t(a)$, thus we get:

$$\mathcal{R}^q_T(\theta^*) = \sum_{t=1}^{T} \mathbb{E}_{a_1,..,a_{t-1}}\big[\mathbb{E}_{a_t\sim q_t}(\mathbb{E}_{a\sim\xi^\theta_t}(\tilde{l}^\theta_t(c_t, a))) - \tilde{l}^\theta_t(c_t, a^*_{c_t})\big].$$

Step 2. Decompose the term $\mathbb{E}_{a\sim\xi^\theta_t}(\tilde{l}^\theta_t(c_t, a))$ in order to use the definition of ξ^θ_t. Indeed, for $\varphi(x) \overset{\text{def}}{=} \frac{1}{\eta^\theta_t}\log \mathbb{E}_{a\sim\xi^\theta_t}\exp(x)$, following the technique described in Section 2.1 of Bubeck (2010), we have:

$$\mathbb{E}_{a\sim\xi^\theta_t}(\tilde{l}^\theta_t(c_t, a)) \;=\; \varphi\big(-\eta^\theta_t(\tilde{l}^\theta_t(c_t, a) - \mathbb{E}_{b\sim\xi^\theta_t}(\tilde{l}^\theta_t(c_t, b)))\big) - \varphi(-\eta^\theta_t\tilde{l}^\theta_t(c_t, a)) \tag{4.5}$$

Now using the fact that $\log x \leqslant x - 1$ and $\exp(-x) - 1 + x \leqslant x^2$, $\forall x \geqslant 0$, the first term on the right hand of (4.5) is bounded by: $\frac{\eta^\theta_t}{2}\mathbb{E}_{a\sim\xi^\theta_t}(\tilde{l}^\theta_t(c_t, a)^2)$

Thus, considering that $\xi^\theta_t(a) = \frac{\exp(-\eta^\theta_t\sum_{s=1}^{t-1}\tilde{l}^\theta_s(c_t,a))}{\sum_a \exp(-\eta^\theta_t\sum_{s=1}^{t-1}\tilde{l}^\theta_s(c_t,a))}$, we can introduce the quantity

$$\Psi^\theta_t(\eta, c) = \frac{1}{\eta}\log\Big(\frac{1}{A}\sum_a \exp\big(-\eta\sum_{s=1}^{t}\tilde{l}^\theta_s(c, a)\big)\Big),$$

so that the second right term of equation(4.5) can be written $\Psi^\theta_{t-1}(\eta^\theta_t, c_t) - \Psi^\theta_t(\eta^\theta_t, c_t)$. Thus we deduce that:

$$\begin{aligned}\mathcal{R}^q_T(\theta^*) \;\leqslant\; \sum_{t=1}^{T}\mathbb{E}_{a_1,..,a_{t-1}}\Big[&\mathbb{E}_{a_t\sim q_t}\big(\frac{\eta^\theta_t}{2}(1 - r_t(a_t))^2\frac{\xi^\theta_t(a_t)}{q^2_t(a_t)}\big)\\ &+\mathbb{E}_{a_t}\big(\Psi^\theta_{t-1}(\eta^\theta_t, c_t) - \Psi^\theta_t(\eta^\theta_t, c_t)\big) - \mathbb{E}_{a_t}\tilde{l}^\theta_t(c_t, a^*_{c_t})\Big],\end{aligned} \tag{4.6}$$

where we have replace $\tilde{l}^\theta_t(c_t, a)$ by its definition.

Step 3. Now we consider the first term in the right hand side of previous equation, which is bounded by:

$$\mathbb{E}_{a_t\sim q_t}((1 - r_t(a_t))^2\frac{\xi^\theta_t(a_t)}{q^2_t(a_t)}) \leqslant \sum_a \frac{\xi^\theta_t(a)}{q_t(a)} \leqslant \frac{1}{\delta}.$$

Step 4. Introduce the equivalence classes. We now consider the second term defined with Ψ functions. Let us introduce the notations $N_t^\theta(c) \stackrel{\text{def}}{=} \sum_{s=1}^{t} \mathbb{I}_{c=[h_{<s}]_\theta}$ for the number of times when the class c is activated up to time t and then $t_c^\theta(i) \stackrel{\text{def}}{=} \min\{t; N_t^\theta(c) = i\}$. Thus we can write:

$$\sum_\theta \sum_{t=1}^{T} (\Psi_{t-1}^\theta(\eta_t^\theta, [h_{<t}]_\theta) - \Psi_t^\theta(\eta_t^\theta, [h_{<t}]_\theta)) = \sum_\theta \sum_{c\in\theta} \sum_{i=1}^{N_T^\theta(c)} \Psi^\theta_{t_c^\theta(i)-1}(\eta^\theta_{t_c^\theta(i)}, c) - \Psi^\theta_{t_c^\theta(i)}(\eta^\theta_{t_c^\theta(i)}, c)$$

$$= \sum_\theta \sum_{c\in\theta} \Big(\sum_{i=1}^{N_T^\theta(c)-1} (\Psi^\theta_{t_c^\theta(i)}(\eta^\theta_{t_c^\theta(i)+1}, c) - \Psi^\theta_{t_c^\theta(i)}(\eta^\theta_{t_c^\theta(i)}, c))\Big) - \Psi^\theta_{t_c^\theta(N_T^\theta(c))}(\eta^\theta_{t_c^\theta(N_T^\theta(c))}, c).$$

Now, by definition of Ψ_t^θ, we also have for any given $a = a_c^{\theta^*}$ (we remind that $\theta = \theta^*$):

$$\begin{aligned} -\Psi^\theta_{t_c^\theta(N_T^\theta(c))}(\eta^\theta_{t_c^\theta(N_T^\theta(c))}, c) &= \frac{\log A}{\eta^\theta_{t_c^\theta(N_T^\theta(c))}} - \frac{1}{\eta^\theta_{t_c^\theta(N_T^\theta(c))}} \log\Big(\frac{1}{A}\sum_a \exp\Big(-\eta^\theta_{t_c^\theta(N_T^\theta(c))} \sum_{s=1}^{t_c^\theta(N_T^\theta(c))} \tilde{l}_s^\theta(c,a)\Big)\Big) \\ &\leqslant \frac{\log A}{\eta^\theta_{t_c^\theta(N_T^\theta(c))}} + \sum_{s=1}^{t_c^\theta(N_T^\theta(c))} \tilde{l}_s^\theta(c,a)\,. \end{aligned}$$

In particular, we can use the optimal action a_c^* when $c = [h_{<t}]_{\theta^*}$.

Step 5. Remark that $\Psi_t^\theta(\cdot, c)$ is increasing for all θ, c. Indeed, we can show that

$$\frac{\partial}{\partial\eta}\Psi^\theta(\eta, c) = \frac{1}{\eta^2} KL(p^\eta_{t,c}, \pi),$$

where π is the uniform distribution over the arms, and $p^\eta_{t,c}(a) = \frac{\exp(-\eta\sum_{s=1}^{t-1}\tilde{l}_s^\theta(c_t,a))}{\sum_a \exp(-\eta\sum_{s=1}^{t-1}\tilde{l}_s^\theta(c_t,a))}$.

Step 6. Now since $\eta^\theta_{t_c^\theta(i)} \leqslant \eta^\theta_{t_c^\theta(i)+1}$, and $\Psi^\theta_{t_c^\theta(i)}(\cdot, c)$ is increasing, we combine the results of each step to deduce that:

$$\mathcal{R}_T^q(\theta^*) \leqslant \mathbb{E}\Big(\sum_c \sum_{i=1}^{N_T^\theta(c)} \frac{\eta^\theta_{t_c^\theta(i)}}{2\delta} + \frac{\log A}{\eta^\theta_{t_c^\theta(N_T^\theta(c))}}\Big).$$

Since we choose $\eta^\theta_{t_c^\theta(i)} = \sqrt{\frac{\delta\log(A)}{i}}$, we get:

$$\mathcal{R}_T^q(\theta^*) \leqslant 2\mathbb{E}\Big(\sum_c \sqrt{\frac{N_T^\theta(c)\log A}{\delta}}\Big) \leqslant 2\sqrt{\frac{T\overline{C}\log A}{\delta}}.$$

□

We now combine Lemma 4.2 and Theorem 4.4 using Exp4 meta algorithm with $\delta = \frac{\gamma}{A}$ to get the final bound:

Theorem 4.5 *For any opponent, the adversarial Φ_Θ-regret of Exp4/Exp3 is bounded as*

$$\tilde{R}_T^\Theta = O(T^{2/3}(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}),$$

where $\overline{C} = \max_{\theta\in\Theta}|\mathcal{H}/\Phi_\theta|$ is the maximum number of classes for models $\theta \in \Theta$.

Proof: Indeed we can apply Theorem 4.4 using Exp4 meta algorithm with $\delta = \frac{\gamma}{A}$. We get:

$$\begin{aligned}\tilde{R}_T^\Theta &= \sum_{t=1}^T \mathbb{E}_{a_1,\dots,a_{t-1}}(r_t(a^*_{[h_{<t}]_{\theta^*}}) - \mathbb{E}_{a_t\sim q_t} r_t(a_t)) \\ &\leqslant \mathcal{R}_T^q(\theta^*) + (e-1)\gamma T + \frac{A\log(|\Theta|)}{\gamma} \\ &\leqslant 2\sqrt{\frac{TA\overline{C}\log A}{\gamma}} + 2\gamma T + \frac{A\log(|\Theta|)}{\gamma}.\end{aligned}$$

We thus choose $\gamma = \frac{(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}}{(4T)^{1/3}}$ to conclude. □

5.1.2 Φ-UCB in the rebel-bandit setting

Similarly, a bound on the Rebel-regret of the Φ^*-UCB algorithm can be derived assuming that we consider a Φ^*-constrained opponent with $\Phi^* = \Phi^{\theta^*} \in \Phi_\Theta$.

Theorem 4.6 (*Rebel regret bound for UCB*) *The Φ_{θ^*}-UCB algorithm in the Rebel bandit setting where $q_t(a) \geqslant \delta$ for all a, and provided $\alpha > 1/2$, satisfies*

$$\mathcal{R}_T^q(\theta^*) \leqslant \sum_{c\in\mathcal{H}/\Phi^*}\sum_{a\neq a_c^*}\Delta_c(a)\Big[\frac{2\alpha\log(T)}{\Delta_c(a)^2\delta} + \sqrt{\frac{\pi\delta\Delta_c(a)^2}{32\alpha\log T}} + c_\alpha\Big]$$

We also have the distribution-free bound:

$$\mathcal{R}_T^q(\theta^*) \leqslant \sqrt{TC^*A}\sqrt{\frac{4\alpha\log(T)}{\delta} + c_\alpha + \sqrt{\frac{\pi\delta}{32\alpha\log(T)}}}.$$

This enables us to deduce the first part of Theorem 4.3, following the same method as Theorem 4.5 but for the stochastic Φ_Θ-regret of Exp4/UCB.

Proof: We write b_t the action proposed by the Φ-UCB algorithm at time t, and a_t the action effectively played according to distribution q_t. We introduce the notations: $N_T(c) = \sum_{t=1}^T \mathbb{I}_{[h_{<t}]=c}$, then $t_c(i) = \min\{t; N_t(c) = i\}$ and for all $a \in \mathcal{A}$, $N_T(c,a) = \sum_{t=1}^T \mathbb{I}_{[h_{<t}]=c}\mathbb{I}_{a_t=a}$. The proof mainly follows the lines of Bubeck (2010). Note that by definition, we want to bound the following term:

$$\mathcal{R}_T^q(\theta^*) = \sum_c\sum_a \Delta_c(a)\mathbb{E}\Big(\sum_{i=1}^{N_T^\theta(c)}\mathbb{I}_{b_t=a}\Big) \tag{4.7}$$

Step one. Decompose the event $b_t = a$. Let us consider a time t for which $[h_{<t}] = c$. Then let us consider a sub-optimal arm a such that $\Delta_c(a) > 0$. Thus it appears that $b_t = a$ if one of the following conditions holds:

(1) $\tilde{\mu}_{t,c}(a_c) \leqslant \mu_c(a_c)$

(2) $\tilde{\mu}_{t,c}(a) > \mu_c(a)$

(3) $\Delta_c(a) < 2\sqrt{\frac{\alpha \log T}{N^\theta_{t-1}(c,a)}}$

Indeed, otherwise we would have

$$\begin{aligned}\tilde{\mu}_c(a_c) &> \mu_c(a_c) = \mu_c(a) + \Delta_c(a)\\ &\geqslant \mu_c(a) + 2\sqrt{\frac{\alpha \log N_t(c)}{N_{t-1}(c,a)}} \geqslant \tilde{\mu}_c(a).\end{aligned}$$

Thus we introduce the quantity $u_c(a) = \frac{4\alpha \log T}{\Delta_c(a)^2}$, and deduce that:

$$\mathbb{E}\Big(\sum_{i=1}^{N^\theta_T(c)} \mathbb{I}_{b_t=a}\Big) \leqslant \mathbb{E}\Big(\sum_{i=1}^{N^\theta_T(c)} \mathbb{I}_{(1)or(2)orN^\theta_{t-1}(c,a)<u_c(a)}\Big).$$

Step 2. Now since $N^\theta_\cdot(c,a)$ is an increasing function of time (note though, that it does not increase by one each time b_t is proposed...), we can define the stopping time $\tau_c(a) = \min\{t; N^\theta_t(c,a) \geqslant u_c(a)\}$, or equivalently the stopping instant $i_c(a) = \min\{i; N^\theta_{t^\theta_c(i)}(c,a) \geqslant u_c(a)\}$. Thus we deduce that:

$$\mathbb{E}\Big(\sum_{i=1}^{N^\theta_T(c)} \mathbb{I}_{b_t=a}\Big) \leqslant \mathbb{E}(i_c(a)) + \mathbb{E}\Big(\sum_{i=i_c(a)+1}^{N^\theta_T(c)} \mathbb{I}_{(1)or(2)}\Big) \tag{4.8}$$

Now we can bound the second term of (4.8) by a constant depending only on α, by an easy peeling argument (we refer to Section 2.2 of Bubeck (2010)):

$$\mathbb{E}\Big(\sum_{i=i_c(a)+1}^{N^\theta_T(c)} \mathbb{I}_{(1)or(2)}\Big) \leqslant 2\mathbb{E}\Big(\sum_{i=i_c(a)+1}^{N^\theta_T(c)} \Big(\frac{\log i}{\log 1/\beta} + 1\Big)\frac{1}{i^{2\beta\alpha}}\Big) \tag{4.9}$$

where $\beta = \frac{1}{\alpha+1/2}$.

Then, we also have, by integration by parts:

$$2\mathbb{E}\Big(\sum_{i=i_c(a)+1}^{N^\theta_T(c)} \Big(\frac{\log i}{\log 1/\beta} + 1\Big)\frac{1}{i^{2\beta\alpha}}\Big) \leqslant 2\int_1^\infty \Big(\frac{\log t}{\log 1/\beta} + 1\Big)\frac{1}{t^{2\beta\alpha}}dt \leqslant \frac{4}{\log(1/\beta)(2\beta\alpha - 1)^2}.$$

Step 3. Thus we focus on the first term $\mathbb{E}(i_c(a))$ of (4.8). Since we know that $q_t(a) \geqslant \delta$ for all a, t, we thus deduce that:

$$\begin{aligned}\mathbb{E}(i_c(a)) &= \sum_{l=0}^{\infty} \mathbb{P}(i_c(a) > l) \leqslant l_0 + \sum_{l=l_0}^{\infty} \mathbb{P}\big(\forall j \leqslant l;\, N^{\theta}_{t^{\theta}_c(j)}(c,a) < u_c(a)\big) \\ &\leqslant l_0 + \sum_{l=l_0}^{\infty} \mathbb{P}\big(\forall j \leqslant l;\, \sum_{i=1}^{j} \mathbb{I}_{a_{t^{\theta}_c(i)}=a} - q_{t^{\theta}_c(i)}(a) < u_c(a) - \delta j\big).\end{aligned}$$

Now by property of martingale difference sequences, we deduce by setting $l_0 = \ulcorner \frac{u_c(a)}{\delta} \urcorner$, that:

$$\begin{aligned}\mathbb{E}(i_c(a)) &\leqslant l_0 + \sum_{l=l_0}^{\infty} \exp(-2(l-l_0)^2\delta^2 l) \\ &\leqslant l_0 + \sum_{l=l_0}^{\infty} \exp(-\frac{(l-l_0)^2}{2\sigma^2}),\end{aligned}$$

where we introduced the quantity $\sigma^2 = \frac{1}{4\delta^2 l_0}$. Thus we deduce that:

$$\mathbb{E}(i_c(a) \leqslant \ulcorner \frac{u_c(a)}{\delta} \urcorner + \sqrt{\frac{\pi}{8}}\sqrt{\frac{\delta}{u_c(a)}}. \tag{4.10}$$

Step 4. Finally, by combining (4.9), (4.10) with (4.8) and (4.7), we deduce the following distribution-dependent bound on the rebel regret:

$$\mathcal{R}^q_T(\theta^*) \leqslant \sum_{c \in \mathcal{H}/\Phi^*} \sum_{a \neq a^*_c} \Delta_c(a) \Big[\frac{2\alpha \log(T)}{\Delta_c(a)^2 \delta} + \sqrt{\frac{\pi\delta\Delta_c(a)^2}{32\alpha \log T}} + c_\alpha\Big],$$

where $c_\alpha = 1 + \frac{4}{\log(\alpha+1/2)}(\frac{\alpha+1/2}{\alpha-1/2})^2$. We deduce the distribution-free bound by the same argument as for Theorem 4.8, remarking that $\sqrt{\frac{\pi}{8}}\sqrt{\frac{\delta\Delta_c(a)^2}{4\alpha \log T}} \leqslant \sqrt{\frac{\pi}{32\alpha \log(T)}} = c'_\alpha$. □

This enables us to deduce the following Theorem, that we prove using the same method as that of Theorem 4.5 but for the stochastic Φ_Θ-regret of Exp4/UCB.

Theorem 4.7 *Assume that we consider a Φ^*-constrained opponent with $\Phi^* \in \Phi_\Theta$, then the stochastic Φ_Θ-regret of Exp4/UCB is bounded as:*

$$R^{\Theta}_T = O\Big((TA)^{2/3}(\overline{C}\log(T))^{1/3}\log(|\Theta|)^{1/2}\Big),$$

where $\overline{C} = |\mathcal{H}/\Phi^*|$ is the number of classes of the model Φ^* of the opponent.

5.2 Playing against an opponent using a known model

In this section, we provide the sanity-check proofs corresponding to the case when the model of the opponent $\varphi^\star$ is known by the learner.

5.2.1 Regret upper bounds against the best history-class-based strategy

Theorem 4.8 *In the case of a Φ-constrained opponent, using the Φ-UCB algorithm with parameter $\alpha > 1/2$, we have the distribution-dependent bound:*

$$R_T^\Phi \leqslant \sum_{c\in\mathcal{H}/\Phi;\mathbb{E}(N_T(c))>0} \sum_{a\in\mathcal{A};\Delta_c(a)>0} \frac{4\alpha\log(T)}{\Delta_c(a)} + \Delta_c(a)c_\alpha$$

where $N_T(c) = \sum_{t=1}^T \mathbb{I}_{[h_{<t}]=c}$, the per-class gaps $\Delta_c(a) \stackrel{\text{def}}{=} \mu_c(a_c^*) - \mu_c(a)$, and the constant $c_\alpha = 1 + \frac{4}{\log(\alpha+1/2)}(\frac{\alpha+1/2}{\alpha-1/2})^2$. We also have a distribution-free bound (i.e. which does not depend on the gaps):

$$R_T^\Phi \leqslant \sqrt{TA\overline{C}\Big(4\alpha\log(T) + c_\alpha\Big)}$$

where $\overline{C} = |\{c \in |\mathcal{H}/\Phi|; \mathbb{E}(N_T(c)) > 0\}|$ is the number of classes that may be activated during the run.

Now, in the case of an arbitrary opponent, using Φ-Exp3 algorithm, we have:

$$\tilde{R}_T^\Phi \leqslant \frac{3}{\sqrt{2}}\sqrt{T\overline{C}A\log(A)}.$$

Proof: **Φ-UCB**: The distribution-dependent bound for Φ-UCB is a direct application of the result of Bubeck (2010) for the algorithm UCB about $\tau_a(t) \stackrel{\text{def}}{=} \sum_{s=1}^t \mathbb{I}_{a_s=a}$ where a_t is played by UCB, that states that $\mathbb{E}(\tau_a(t)) \leqslant \frac{4\alpha\log(t)}{\Delta_c^2(a)} + c_\alpha$. Indeed, we use the fact that $R_T^\Phi = \sum_{c\in\mathcal{H}/\Phi} R_T(c)$ and thus remark that when a class c is visited, then we play according to a UCB algorithm for this class.

Thus, for the distribution-free bound, we have:

$$\begin{aligned}
R_T^\Phi &= \sum_c \sum_a \Delta_c(a)\mathbb{E}(\tau_a(N_T(c))) \\
&\leqslant \sum_c \sum_a \sqrt{\mathbb{E}(\tau_a(N_T(c)))}\sqrt{4\alpha\log(T) + c_\alpha} \\
&\leqslant \sum_c \sqrt{\mathbb{E}(N_T(c))}\sqrt{A}\sqrt{4\alpha\log(T) + c_\alpha} \\
&\leqslant \sqrt{T\overline{C}A}\sqrt{4\alpha\log(T) + c_\alpha},
\end{aligned}$$

where we used that $\sum_a \tau_a(s) = s$ for all s, and $\sum_c N_T(c) = T$, and the Cauchy-Schwartz inequality twice.

Φ-Exp3: The bound for Φ-Exp3 follows from the bound of the anytime version of the Exp3 algorithm. Indeed we have

$$\tilde{R}_T^\Phi \leqslant \sum_c \mathbb{E}\Big(\frac{A}{2}\sum_{i=1}^{N_T(c)} \eta_i^c + \frac{\log(A)}{\eta^c_{N_T(c)}}\Big)\,,$$

we deduce the bound by setting $\eta_i^c = \sqrt{\frac{2\log A}{Ai}}$. □

5.2.2 Lower bounds on the regret

Theorem 4.9 *Let* sup *represents the supremum taken over all* Φ*-constrained opponents and* inf *the infimum over all forecasters, then the stochastic* Φ*-regret is lower-bounded as:*

$$\sup_{\Phi;|\mathcal{H}/\Phi|=C} \inf_{algo} \sup_{\Phi-opp} R_T^\Phi \geqslant \frac{1}{20}\sqrt{TAC}.$$

Now, let sup *represents the supremum taken over all possible opponents and* inf *the infimum over all forecasters, then the adversarial* Φ*-regret is lower-bounded as:*

$$\sup_{\Phi;|\mathcal{H}/\Phi|=C} \inf_{algo} \sup_{opp} \tilde{R}_T^\Phi \geqslant \frac{1}{20}\sqrt{TAC}.$$

Proof: Let us fix the horizon T and the number of classes C. We consider the opponent defined using the specific class-function Φ such that each class c is periodically visited every C time steps, thus T/C times. Note that $T = \frac{T}{C}C$ and that this is intuitively the opponent that makes the algorithm switch between classes the most.

Now we define more precisely the rewards output by the opponent. Let us consider the stochastic bandits such that for each class c, one arm a_c is a Bernoulli $B((1+\varepsilon_c)/2)$, and all others are $B((1-\varepsilon_c)/2)$.

Then by application of Lemma 2.2 in Bubeck (2010), for ε_c of order $\sqrt{\frac{A}{s}}$, we have in the Bandit information setting the following inequality:

$$\sup_{\{a_c\}_c} \sum_{t=1}^{s} \mu_c(a_c) - \mu_c(a_t) \geqslant s\varepsilon_c\Big(1 - \frac{1}{A} - \sqrt{\frac{s\varepsilon_c}{2A}\log(\frac{1+\varepsilon_c}{1-\varepsilon_c})}\Big).$$

Thus with the notations $N_T(c) \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{c=[h_{<t}]}$ for the number of times when class c is activated up to time t and then $t_c(i) \overset{\text{def}}{=} \min\{t; N_t(c) \geqslant i\}$, we deduce that:

$$\begin{aligned}\sup_{\{a_c\}_c} \sum_c \sum_{t=1}^{T} (\mu_c(a_c) - \mu_c(a_t))\mathbb{I}_{c=[h_{<t}]} &= \sum_c \sup_{a_c} \sum_{i=1}^{N_T(c)} (\mu_c(a_c) - \mu_c(a_{t_c(i)})) \\ &\geqslant \sum_c N_T(c)\varepsilon_c\Big(1 - \frac{1}{A} - \sqrt{\varepsilon_c \log(\frac{1+\varepsilon_c}{1-\varepsilon_c})}\sqrt{\frac{N_T(c)}{2A}}\Big).\end{aligned}$$

Since the a_c are chosen by the opponent such that each class is visited exactly $N_T(c) = T/C$ times, then we deduce that the Φ-pseudo-regret is lower-bounded as:

$$\sup_{\{a_c\}_c} \sum_c \sum_{t=1}^{T} (\mu_c(a_c) - \mu_c(a_t)) \mathbb{I}_{c=[h_{<t}]} \geqslant \sum_c \frac{T}{C} \varepsilon_c \Big(1 - \frac{1}{A} - \sqrt{\varepsilon_c \log(\frac{1+\varepsilon_c}{1-\varepsilon_c})} \sqrt{\frac{T}{2AC}} \Big).$$

Thus, after some tedious computations to optimize ε_c, we finally get a lower bound of order: $\frac{1}{20} \sum_c \sqrt{\frac{T}{C} A} = \frac{1}{20} \sqrt{TAC}$. Note that this is valid only if $\varepsilon_c \sim \sqrt{\frac{A}{N_T(c)}}$ is small (less than 1), i.e. if the number of classes C is smaller than a constant times $\frac{T}{A}$ (and if this not the case, the lower bound becomes obviously of order T).

The second part of the Theorem can be proved using the same construction. □

5.3 Approximation error of the models

The following result sheds light on a specific term that appears to be an approximation term of the true model θ^* by other models θ.

Theorem 4.10 (*Approximation error of bandit models*) *For any $(p_t(\theta))_{t,\theta} \in [0,1]$, thus for any meta algorithm run on top of Exp3 algorithm and defined with $q_t(a) = \sum_\theta p_t(\theta) \xi_t^\theta(a)$ and decreasing coefficient η_t^θ, the following holds true:*

$$\begin{aligned} \tilde{R}_T^\Theta \;\leqslant\; & \mathbb{E}\Big(\sum_\theta \sum_{c \in \theta} \sum_{i=1}^{N_T^\theta(c)} \frac{\eta^\theta_{t_c^\theta(i)} A}{2} p_{t_c^\theta(i)}(\theta) + \sum_\theta \sum_{c \in \theta} \frac{\log A}{\eta^\theta_{t_c^\theta(N_T^\theta(c))}} \\ & + \sum_\theta \sum_{c \in \theta} \inf_{a_c^\theta} \sum_{i=1}^{N_T^\theta(c)} (r_{t_c^\theta(i)}(a^*_{[h_{<t_c^\theta(i)}]}) - r_{t_c^\theta(i)}(a_c^\theta)) p_{t_c^\theta(i)}(\theta) \Big). \end{aligned}$$

The term on the second line is actually a mixture of approximation errors of each model, and it seems it can not be reduced without further assumption on the quality of the models.

Proof: The proof is in four steps.

Step 1. Rewrite the regret to make appear the probabilities $\xi_t^\theta(a)$. We first introduce:

$$\begin{aligned} R_T &= \sum_{t=1}^{T} r_t(a_{[h_{<t}]_*}) - r_t(a_t) \\ &= \sum_\theta \sum_{t=1}^{T} \mathbb{E}_{a_t \sim q_t} \Big(\frac{r_t(a_t) p_t(\theta)}{q_t(a_t)} \mathbb{I}_{a_t = a_{[h_{<t}]_*}} \Big) - \frac{r_t(a_t) \xi_t^\theta(a_t) p_t(\theta)}{q_t(a_t)}. \end{aligned}$$

Now we have: $\tilde{l}_t^\theta(c_\theta, a) = (1 - r_t(a))\frac{p_t(\theta)}{q_t(a)}\mathbb{I}_{a_t=a}\mathbb{I}_{c_\theta=[h_{<t}]_\theta}$, thus taking the expectation over a_t for each time t, we have:

$$\begin{aligned}
\tilde{R}_T &= \sum_{t=1}^{T} \mathbb{E}_{a_r}(r_t(a_{[h_{<t}]_*} - r_t(a_t)) \\
&= \sum_{\theta}\sum_{t=1}^{T} \mathbb{E}_{a_t}\Big(\frac{p_t(\theta)}{q_t(a_t)}\mathbb{I}_{a_t=a_{[h_{<t}]_*}} - \tilde{l}_t^\theta([h_{<t}]_\theta, a_{[h_{<t}]_*}))\Big) \\
&\quad + \sum_{\theta}\sum_{t=1}^{T} \mathbb{E}_{a_t}\Big(\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}_t^\theta([h_{<t}]_\theta, a)) - \frac{p_t(\theta)\xi_t^\theta(a_t)}{q_t(a_t)}\Big).
\end{aligned}$$

We can simplify the above expression since $\mathbb{E}_{a_t}(\frac{p_t(\theta)}{q_t(a_t)}\mathbb{I}_{a_t=a_{[h_{<t}]_*}}) = \mathbb{E}_{a_t}(\frac{p_t(\theta)\xi_t^\theta(a_t)}{q_t(a_t)}) = p_t(\theta)$.

Step 2. Decompose the term $\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}_t^\theta([h_{<t}]_\theta, a))$ in order to use the definition of ξ_t^θ. Indeed, one can upper bound this term by

$$\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}_t^\theta([h_{<t}]_\theta, a)) \leqslant \frac{\eta_t^\theta}{2}\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}_t^\theta([h_{<t}]_\theta, a)^2) - \frac{1}{\eta_t^\theta}\log(\sum_a \exp(-\eta_t^\theta \tilde{l}_t^\theta([h_{<t}]_\theta, a)\xi_t^\theta(a))).$$

Thus, since by definition we have that $\xi_t^\theta(a) = \frac{\exp(-\eta_t^\theta \sum_{s=1}^{t-1}\tilde{l}_s^\theta([h_{<t}]_\theta, a))}{\sum_a \exp(-\eta_t^\theta \sum_{s=1}^{t-1}\tilde{l}_s^\theta([h_{<t}]_\theta, a)}$, we can introduce the quantity $\Psi_t^\theta(\eta, c) = \frac{1}{\eta}\log(\frac{1}{A}\sum_a \exp(-\eta\sum_{s=1}^{t}\tilde{l}_s^\theta(c, a)))$ so that the previous regret term writes:

$$\begin{aligned}
\tilde{R}_T &\leqslant \sum_{\theta}\sum_{t=1}^{T} \mathbb{E}_{a_t}(\frac{\eta_t^\theta}{2}(1 - r_t(a_t))^2\frac{p_t^2(\theta)\xi_t^\theta(a_t)}{q_t^2(a_t)}) \\
&\quad + \sum_{\theta}\Big(\sum_{t=1}^{T} \mathbb{E}_{a_t}(\Psi_{t-1}^\theta(\eta_t^\theta, [h_{<t}]_\theta) - \Psi_t^\theta(\eta_t^\theta, [h_{<t}]_\theta)) - \mathbb{E}_{a_t}(\tilde{l}_t^\theta([h_{<t}]_\theta, a_{[h_{<t}]_*}))\Big).
\end{aligned}$$

Step 3. Introduce the equivalence classes. We now consider the term in the right hand side of the above equation defined with Ψ functions. Note that we do not change the bound on the term $\tilde{R}_T$ by considering the sum over the θ such that $p_t(\theta) > 0$.

Let us introduce the following notations $N_t^\theta(c) = \sum_{s=1}^{t}\mathbb{I}_{c=[h_{<s}]_\theta}\mathbb{I}_{p_s(\theta)>0}$ and then similarly $t_c^\theta(i) = \min\{t; N_t^\theta(c) = i\}$. Thus we can write:

$$\begin{aligned}
&\sum_\theta \sum_{t=1}^{T}(\Psi_{t-1}^\theta(\eta_t^\theta, [h_{<t}]_\theta) - \Psi_t^\theta(\eta_t^\theta, [h_{<t}]_\theta))\mathbb{I}_{p_t(\theta)>0} \\
= &\sum_\theta \sum_{c\in\theta}\sum_{i=1}^{N_T^\theta(c)} \Psi_{t_c^\theta(i)-1}^\theta(\eta_{t_c^\theta(i)}^\theta, c) - \Psi_{t_c^\theta(i)}^\theta(\eta_{t_c^\theta(i)}^\theta, c) \\
= &\sum_\theta \sum_{c\in\theta}\sum_{i=1}^{N_T^\theta(c)-1}(\Psi_{t_c^\theta(i)}^\theta(\eta_{t_c^\theta(i)+1}^\theta, c) - \Psi_{t_c^\theta(i)}^\theta(\eta_{t_c^\theta(i)}^\theta, c) \\
&\qquad - \Psi_{t_c^\theta(N_T^\theta(c))}^\theta(\eta_{t_c^\theta(N_T^\theta(c))}^\theta, c).
\end{aligned}$$

Now, by definition of Ψ_t^θ, for any given $a = a_c^\theta$ we have

$$\begin{aligned} -\Psi^\theta_{t_c^\theta(N_T^\theta(c))}(\eta^\theta_{t_c^\theta(N_T^\theta(c))}, c) &= \frac{\log A}{\eta} - \frac{1}{\eta}\log\Big(\frac{1}{A}\sum_a \exp\big(-\eta \sum_{s=1}^{t_c^\theta(N_T^\theta(c))} \tilde{l}_s^\theta(c,a)\big)\Big) \\ &\leqslant \frac{\log A}{\eta} + \sum_{s=1}^{t_c^\theta(N_T^\theta(c))} \tilde{l}_s^\theta(c,a)\,, \end{aligned}$$

where η is a shorthand notation for $\eta^\theta_{t_c^\theta(N_T^\theta(c))}$.

Step 4. Now since $\eta^\theta_{t_c^\theta(i)} \leqslant \eta^\theta_{t_c^\theta(i)+1}$ and $\Psi^\theta_{t_c^\theta(i)}(\cdot, c)$ is increasing for all θ, c, we deduce from the previous equations that:

$$\begin{aligned} \tilde{R}_T \leqslant & \sum_\theta \sum_{c\in\theta} \sum_{i=1}^{N_T^\theta(c)} \sum_a \frac{\eta^\theta_{t_c^\theta(i)}}{2} \frac{p^2_{t_c^\theta(i)}(\theta)\xi^\theta_{t_c^\theta(i)}(a)}{q_{t_c^\theta(i)}(a)} + \sum_\theta \sum_{c\in\theta} \frac{\log A}{\eta^\theta_{t_c^\theta(N_T^\theta(c))}} \\ & + \sum_\theta \sum_{c\in\theta} \inf_{a_c^\theta} \sum_{t=1}^{t_c^\theta(N_T^\theta(c))} \mathbb{E}_{a_t}(\tilde{l}_t^\theta(c, a_c^\theta) - \tilde{l}_t^\theta(c, a_{[h_{<t}]_*})). \end{aligned}$$

Now we conclude by taking the expectation, seeing that $p_{t_c^\theta(i)}(\theta)\xi_{t_c^\theta(i)} \leqslant q_{t_c^\theta(i)}(a)$, and that by definition $\mathbb{E}_{a_t}(\tilde{l}_t^\theta(c, a_c^\theta)) = (1 - r_t(a_c^\theta))p_t(\theta)\mathbb{I}_{c=[h_{<t}]_\theta}$. □

Part II

The Batch World: Randomization and Sampling.

After this first part where we studied some variations of the bandit problem, which can be seen as a pure online problem when compared to the general reinforcement learning problem, we now turn to the study of some important questions that concern so-called batch learning, i.e. when we are given a batch of data instead of a stream of data.

Why studying batch learning is important for sequential learning. First studying batch learning is interesting by itself, actually almost all Machine Learning is about batch data, and it involves important notions such as regression, sampling or learning complexity. This is also useful for the study of the sequential learning setting, which is maybe less obvious. The general reason here is that since learning with a given batch data is easier than learning while sequentially acquiring data - the notion of regret actually measures the gap of performance - then it makes sense to first study the fundamental questions of batch learning before developing similar tools for the sequential learning problem. Note also that studying batch learning is a priori not sufficient since major questions such as for instance the exploration-exploitation tradeoff, which is fundamental in sequential learning, do not appear in batch learning.

One example of transfer of knowledge from batch to sequential learning is that the precise understanding of a regression problem with random design is a natural first step before addressing one challenging task of reinforcement learning that makes use of regression with Markov design, see Lazaric et al. (2010a). Another includes an extension of concentration inequalities to their self-normalized version, that we showed to be useful in the bandit setting, see chapter 2, and a recent successful idea considers empirical complexity for batch data such as the Rademacher complexity that has been extended to the case of sequential learning, see Rakhlin et al. (2010), in order to derive a control on the empirical process similar to batch theory.

Now, there are of course plenty of important concepts coming from statistical theory for batch data that still need to be understood from a sequential point of view - data driven penalties, local CLT, Sanov's theorem, PAC analysis, transport - and this opens a wide avenue of research in order to develop the corresponding concepts for non-asymptotic theory of sequential learning.

Contributions. In Chapter 5, we present a general survey of tools coming from statistical theory that we gather here for the sake of clarity, since most of the theorems presented in this chapter are used here and there in this Ph.D. dissertation.

In Chapter 6, we are interested in the use of random matrices in a setting of regression with random design. Interestingly, if the tools needed to assess performance bounds of the proposed estimators have only been popularized quite recently due to the many theoretical and practical developments that the topic of random matrices has lead to in the last few years, we realized that such idea of using random projections, or random representations as they are called in Sutton (1996), have been used for long in more applied settings, like robotics, or texture synthesis for instance, which gives additional motivation for understanding them.

For instance, Richard Sutton already studied experimentally the effect of randomization in neural networks in Sutton and Whitehead (1993) and also highlighted that the 1962 Rosenblatt's perceptron was originally used with a pre-randomization layer in order to improve performance.

In Chapter 7, we consider another use of random matrices in a more traditional way linked with the problem of recovery of a sparse function. Specifically, we show how the use of random integral operators enables to relax the traditional assumptions of (almost) orthogonality of the underlying dictionary, turning the recovery problem into a simple integration problem.

Finally in Chapter 8, although slightly disconnected with the rest of this manuscript, we analyze the Rademacher complexity of a problem known as multi-view learning.

Chapter 5

Statistical Learning.

In this chapter, we present theorems and results coming from different areas of statistical learning theory that we either use in some chapters of this thesis or that we just consider to be important for the current and future development of bandit and reinforcement learning theory. We think it is more convenient to have such theorems and pointers gathered in a dedicated chapter rather than spread here and there.

Contents

1 The concentration of measure phenomenon

The concentration of measure phenomenon Ledoux (2001), and its implementation in terms of concentration inequalities is certainly the most useful tool of the last decade in order to prove performance bounds of algorithms in machine learning. It enables to derive results not only asymptotically, but also for a finite amount of data.

We present in this section different types of inequalities. First we consider concentration inequalities for a single random identically distributed (i.i.d.) process, then we consider more refined results useful for decision processes, which include results for martingales. We finally quickly present some results about random matrices.

1.1 Concentration inequalities for i.i.d. sequences

Hoeffding's inequality The following simple Lemma relates the logarithm of the moment generating function of a bounded random variable to its expectation (see Hoeffding (1963)).

Lemma 5.1 (Hoeffding's lemma) *Let X be real random variable that is almost surely contained in the interval $[a, b]$. Then for all $\lambda \in \mathbb{R}$,*

$$\log(\mathbb{E}(\exp(\lambda X))) \leqslant \lambda \mathbb{E}(X) + \frac{\lambda^2 (b-a)^2}{8}$$

This lemma is at the core of many useful results, like the better known inequality from Hoeffding (1963), that can be seen as a non asymptotic law of large numbers.

Theorem 5.1 (*Hoeffding's inequality*) *Let $\{X_i\}_{1 \leqslant i \leqslant n}$ be real centered independent random variables, such that for all $i \leqslant n$, $X_i \in [a_i, b_i]$ almost surely. Then for any $\varepsilon \geqslant 0$, we have*

$$\mathbb{P}(\sum_{i=1}^{n} X_i \geqslant \varepsilon) \leqslant \exp\Big(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big)$$

This theorem is of practical interest since it gives a bound of the mean of random variables that only depends on the support of the unknown law, which explains the tremendous amount of results that directly use this inequality. Note thus, that it is natural to ask to which extent the bound provided by Hoeffding's inequality is tight. For instance, the following Bernstein's inequality (approximation of Bennett's inequality) may be applied when additional information is known about the variance of the law, and is generally tighter than Hoeffding when the variance is small, see Bernstein (1924).

Theorem 5.2 (*Bernstein's inequality*) *Let $\{X_i\}_{1 \leqslant i \leqslant n}$ be independent real valued random variables and assume that there exist two positive numbers v_n and d_n such that:*

$$\sum_{i=1}^{n} \mathbb{E}(X_i^2) \leqslant v_n \text{ and } \forall r \in \mathbb{N}; r \geqslant 3, \ \sum_{i=1}^{n} \mathbb{E}[(X_i)_+^r] \leqslant \frac{r!}{2} v_n d_n^{r-2}\,.$$

Let $S_n \overset{\text{def}}{=} \sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))$, then for any $x \geqslant 0$, we have

$$\mathbb{P}(S_n \geqslant \sqrt{2v_n x} + d_n x) \leqslant \exp(-x).$$

Note also that historically, other inequalities was proved by Sergei Bernstein, and then were rediscovered several times in various forms. Thus, Chernoff's bound, Hoeffding's inequality and Azuma's inequality are in fact special cases of original Bernstein's inequalities.

Orlicz norms A convenient way to understand the previous result, is to consider that the knowledge of the variance gives control on the tails of the distribution of the random variable. More generally, such a control can be understood by means of Orlicz spaces and Orlicz' norms.

Definition 5.1 (*Orlicz' space*) *A Young-Orlicz modulus is a convex increasing function ψ from $[0,\infty)$ onto $[0,\infty)$ (thus $\psi(0)=0$ and $\psi(x)\to\infty$ when $x\to\infty$). Let (X,τ,μ) be a measure space and ψ be a Young-Orlicz modulus. Denote by $\mathcal{L}_\psi(X,\tau,\mu)$ the space of all real-valued measurable functions f onto X such that*

$$||f||_\psi := \inf\{c>0 : \mathbb{E}_\mu(\psi(|f(X)|/c)) \leqslant 1\} < \infty.$$

Then $L_\psi(X,\tau,\mu)$, the set of all the equivalence classes of functions in $\mathcal{L}_\psi(X,\tau,\mu)$ for the almost everywhere equality, is called an Orlicz' space.

Note that there are others equivalent definitions of Orlicz norms, see Pollard (1990). A specific Young-Orlicz modulus is the following one $\psi_\alpha(x) := \exp(x^\alpha) - 1$, $\alpha \geqslant 1$. Note that we have $L_\infty \subset L_{\psi'_\alpha} \subset L_{\psi_\alpha} \subset L_2$ for all α, $\alpha' \leqslant \alpha$. This modulus has the following interesting property (see Pollard (1990)) that is directly related to a control on the tail probabilities:

Proposition 5.1 *Let X be a real-valued random variable. Then the following statements are equivalent:*

- $\exists c>0; ||X||_{\psi_\alpha} \leqslant c$.
- $\exists c_1, c_2; \forall t>0, \quad \mathbb{P}(|X| \geqslant t) \leqslant c_1 \exp\big(-(\frac{t}{c_2})^\alpha\big)$.

Moreover if $||X||_{\psi_\alpha} \leqslant c$, then the second line holds with $c_1 = 2$ and $c_2 = c$.

Indeed such a proposition shows that controlling a ψ_α-Orlicz norm provides a very precise behavior on the deviation of X, and thus enables to get concentration results for sums of random variables more precise than Bernstein type inequalities. This result is actually more general, see Pollard (1990) for very nice results around this notion.

Sanov's theorem Now, in order to even better understand the tightness of the Hoeffding's inequality, one may want to get bounds that depend on the law of the random variable itself, and not only on features such as its support, its variance or its α-orlicz norm. Such a result has been developed for instance in Sanov (1957) and is known as Sanov's Theorem. See also Csiszár (1984), Dembo and Zeitouni (1998) for important refinements. First, the following asymptotic statement gives the precise distribution-dependent quantity that should appear in the previous inequalities. Let us write ν a distribution probability and $\widehat{\nu}_n$ be the empirical distribution defined using n samples i.i.d. from ν.

Theorem 5.3 (*Sanov's theorem*) *Let $\mathcal{D}$ be a set of distributions over a space $\mathcal{X}$, endowed with the topology of narrow convergence. We have*

$$\liminf_n \frac{1}{n} \log \mathbb{P}(\widehat{\nu}_n \in \mathring{\mathcal{D}}) \geqslant -\Lambda_\nu(\mathring{\mathcal{D}})$$
$$\limsup_n \frac{1}{n} \log \mathbb{P}(\widehat{\nu}_n \in \bar{\mathcal{D}}) \leqslant -\Lambda_\nu(\bar{\mathcal{D}})$$

where $\Lambda_\nu(A) \stackrel{\text{def}}{=} \inf_{\nu' \in A} \mathcal{K}(\nu', \nu)$ and $\mathcal{K}(\nu', \nu)$ is the Kullback-Leibler divergence between the distributions ν' and ν.

Non-asymptotic behavior In order to relate this quantity to the Hoeffding's bound, we need the non-asymptotic behavior counterpart of the previous inequality. In the case of convex set of distributions, we have this result by Dinwoodie (1992). See also Csiszár (1984) for a more general result on so-called *almost completely* convex sets.

Theorem 5.4 (*Non asymptotic Sanov's theorem for convex sets*) *Let $\mathcal{D}$ be a convex set of distributions over a space $\mathcal{X}$. Then, we have, for all $n \geqslant 0$,*

$$\mathbb{P}(\widehat{\nu}_n \in \mathring{\mathcal{D}}) \leqslant \exp(-n\Lambda_\nu(\mathring{\mathcal{D}})) \,.$$

Unfortunately, this result do not generalize nicely to the case of non-convex sets of distributions. In the general case, without any convexity assumption on $\mathcal{D} \subset \mathcal{M}([0,1])$, we only have the following result, specified from Dembo and Zeitouni (1998) for distributions with support included in $[0,1]$.

Theorem 5.5 (*General non asymptotic Sanov's theorem*) *We have, for all $n \geqslant 0$,*

$$\mathbb{P}(\widehat{\nu}_n \in \mathcal{D}) \leqslant \inf \left\{ M(\mathcal{M}([0,1]), \delta) e^{-n\Lambda_\nu(\mathcal{D}^\delta)} \,;\, \delta > 0 \right\}$$

where $\mathcal{M}([0,1])$ is the set of probability measures on $[0,1]$, $M(A,\delta)$ is the minimal number of balls of radius δ that covers the set A and $\mathcal{D}^\delta = \{\nu' \in \mathcal{M}([0,1]); d_{levy}(\nu', \mathcal{D}) \geqslant \delta\}$ is the enlargement of the the set $\mathcal{D}$ for the levy distance.

Pinsker's inequality Sanov's theorem gives the precise asymptotic behavior that should appear in the case of Hoeffding's inequality. This behavior makes appear the Kullback-Leibler divergence as a fundamental underlying quantity while Hoeffding's inequality only considers the mean; let us focus on a distribution ν with support included in $[0,1]$, mean μ and empirical mean $\widehat{\mu}_n$ and distribution $\widehat{\nu}_n$. We have on the one hand by Hoeffding's inequality, for all $\delta > 0$

$$\mathbb{P}(\widehat{\mu}_t \geqslant \mu + \delta) \leqslant \exp(-t2\delta^2),$$

and on the other hand by the non-asymptotic Sanov's theorem

$$\mathbb{P}(\widehat{\nu}_t \in \mathcal{D}_+(\mu+\delta)) \leqslant \exp(-t\Lambda_\nu(\mathcal{D}_+(\mu+\delta)))\,,$$

where $\mathcal{D}_+(x) \stackrel{\text{def}}{=} \{\nu'; \nu' \text{ has mean } \mu' \text{ higher than } x\}$ is the convex set defined such that

$$\mathbb{P}(\widehat{\mu}_t \geqslant \mu+\delta) = \mathbb{P}(\widehat{\nu}_t \in \mathcal{D}_+(\mu+\delta))\,.$$

In order to link $2\delta^2$ to the quantity $\Lambda_\nu(\mathcal{D}_+(\mu+\delta))$ that makes use of the Kullback-Leibler divergence, and thus understand how tight is Hoeffding's inequality, we now we make use of Pinsker's inequality (see e.g. Cover and Thomas (1991)).

Lemma 5.2 (Pinsker's inequality) *For all distributions ν, ν', then*

$$\mathcal{K}(\nu,\nu') \geqslant \frac{||\nu-\nu'||_{TV}^2}{2},$$

where $||\nu-\nu'||_{TV}$ is the total variation norm between ν and ν'.

Note that there is a real gap between the two quantities involved in Pinsker's inequality. Indeed let us mention the following improved bound, due to Fedotov et al. (2003)

$$KL(\nu||\nu') \geqslant \frac{||\nu-\nu'||_{TV}^2}{2} + \frac{||\nu-\nu'||_{TV}^4}{36} + \frac{||\nu-\nu'||_{TV}^6}{270} + \sum_{k=4}^{24} c_k ||\nu-\nu'||_{TV}^{2k},$$

where all the c_k are positive explicit optimal terms, i.e. we can not extend this bound to higher order polynomials with all positive monomes.

Now in our case we simply combine the above inequality together with the following simple remark for distributions with support in $[0,1]$.

Lemma 5.3 *In the case when ν and ν' have support included in $[0,1]$, then we have the inequality $||\nu-\nu'||_{TV} \geqslant 2|\mu-\mu'|$.*

Proof: Indeed, it can be shown that $||\nu-\nu'||_{TV} = 2(\nu(A)-\nu'(A))$, where $A = \{x : \frac{\nu(x)}{\nu'(x)} \geqslant 1\}$. Then, since $x \in [0,1]$, we also have that $\nu(A)-\nu'(A) = \int_A \nu(x)-\nu'(x) \geqslant \int_A x(\nu(x)-\nu'(x)) + \int_{\overline{A}} x(\nu(x)-\nu'(x))$, where the second term is negative by definition of A. Thus we deduce that $||\nu-\nu'||_{TV} \geqslant 2(\mu-\mu')$, and we conclude by symmetry. □

Note that in general, for a given set of distributions $\mathcal{D}$, lower bounding the term $\Lambda_\nu(\mathcal{D})$ may be a difficult task. Let us refer the interested reader to some useful insights coming from Geometrical information theory (see Amari and Nagaoka (2000)), as well as from Transport theory (see Gozlan and Léonard (2010)) that precisely addresses such questions, as well as

many other extensions. In our case where $\mathcal{D} = \mathcal{D}_+(\mu + \delta)$, a lower bound simply follows by the previous inequalities:

$$\begin{aligned}\Lambda_\nu(\mathcal{D}_+(\mu+\delta)) &= \inf\{KL(\nu'||\nu);\ \nu' \text{ s.t. } \mu' \geqslant \mu+\delta\} \\ &\geqslant \inf\{\frac{(2|\mu'-\mu|)^2}{2};\ \nu' \text{ s.t. } \mu' \geqslant \mu+\delta\} \\ &\geqslant 2\delta^2 + \frac{4\delta^4}{9} + \frac{2^6\delta^6}{270} + \sum_{k=4}^{24} c_k(2\delta)^{2k} > 2\delta^2\,,\end{aligned}$$

which shows more precisely in which sense we can say that the Hoeffding's inequality is not tight.

1.2 Concentration inequalities for decision processes

Maximal and self-normalized inequalities Concentration inequalities are nice but we may often want a more uniform result over the number of samples. Instead of using a naive union bound that is generally suboptimal, we can instead resort to so-called maximal inequalities. For instance the following bound is derived from Azuma's inequality for martingales, see Azuma (1967).

Theorem 5.6 (*Maximal Hoeffding's inequality*) *Let $\{X_i\}_{1\leqslant i\geqslant n}$ be a sequence of non-negative independent random variables $X_i \in [0,1]$, $\mathbb{P}$-a.s. with mean μ. Then for all $x > 0$,*

$$\mathbb{P}(\sup_{1\leqslant k\leqslant n} \frac{\sum_{i=1}^k X_i - k}{\sqrt{n}} \geqslant x) \leqslant e^{-2x^2}\,.$$

Note that here, the term after the supremum is normalized by $\sqrt{n}$ whereas it would seem more natural to have a normalization by a factor $\sqrt{k}$ instead. This is exactly the purpose of self-normalized inequalities, which can be seen as a stronger notion. In Garivier and Leonardi (2010), the author proposed such self-normalized inequalities, we only report one simplified version here.

Let F_n the σ-field of the past (so that for $k > n$, X_k is independent from F_n) and let us consider a predictable sequence $(\varepsilon_i)_{i\geqslant 1}$ of Bernoulli variables (i.e. such that for all $i > 0$, ε_i is F_{i-1}-measurable).

We further introduce the following notations:

$$S_k = \sum_{i=1}^k X_i\varepsilon_i,\ \ M_k = \sum_{i=1}^k \mu_i\varepsilon_i,\ \ N_k = \sum_{i=1}^k \varepsilon_i\,.$$

Theorem 5.7 (*Self-Normalized Hoeffding's inequality*) *For all $x > 0$, for all $\eta > 0$,*

$$\mathbb{P}(\sup_{1\leqslant k\leqslant n} \frac{S_k - M_k}{\sqrt{N_k}} \geqslant x) \leqslant \ulcorner \frac{\log(n)}{\log(1+\eta)} \urcorner e^{-2x^2(1-\frac{\eta^2}{16})}\,.$$

The idea of the proof is to consider a cylinder of high probability around the trajectory of the empirical process, which enables to replace the naive union bound over all possible instants by a peeling argument, and thus get a rate similar to the non-maximal inequality up to a logarithmic factor. Other tight results can be found in Garivier (2011) and are related to the law of the Iterated logarithm.

Interestingly, another consequence of the analysis developed in Garivier and Leonardi (2010) is that one can prove the following inequality in the case of Bernoulli random variables

Lemma 5.4 *For all $p \in [0,1]$, all $\varepsilon > 1$, and all $k \geqslant 1$,*

$$\mathbb{P}_p\left\{ N_k \, \mathcal{K}\Big(\beta\big(\widehat{p}_{N_k}\big), \, \beta(p)\Big) \geqslant \varepsilon \right\} \leqslant 2e \left\lceil \varepsilon \log k \right\rceil e^{-\varepsilon} .$$

where $\widehat{p}_{N_k} = \frac{S_k}{N_k}$ and $\beta(p)$ is the law of a Bernoulli random variable with parameter p.

Concentration inequalities for martingales We now detail some important results for concentration of martingales sequences.

Following the important work of Bercu and Touati (2008), where some refinements of the following results can be found, let us consider (M_n), a locally square integrable real martingale adapted to a filtration $F = (F_n)$ with $M_0 = 0$. The predictable quadratic variation and the total quadratic variation of (M_n) are respectively given by

$$\langle M \rangle_n = \sum_{k=1}^{n} \mathbb{E}[\Delta M_k^2 | F_{k-1}] \quad \text{and} \quad [M]_n = \sum_{k=1}^{n} \Delta M_k^2$$

where $\Delta M_n = M_n - M_{n-1}$. The celebrated Azuma–Hoeffding inequality is as follows.

Theorem 5.8 (*Azuma–Hoeffding's inequality*) *Let (M_n) be a locally square integrable real martingale such that, for each $1 \leqslant k \leqslant n$, $a_k \leqslant \Delta M_k \leqslant b_k$ a.s. for some constants $a_k < b_k$. Then, for all $x \geqslant 0$,*

$$P(|M_n| \geqslant x) \leqslant 2 \exp\Big(- \frac{2x^2}{\sum_{k=1}^{n} (b_k - a_k)^2} \Big) .$$

The next result is from (Freedman, 1975, Th. 1.6).

Theorem 5.9 (*Freedman's inequality*) *Let (M_n) be a locally square integrable real martingale such that, for each $1 \leqslant k \leqslant n$, $|\Delta M_k| \leqslant c$ a.s. for some constant $c > 0$. Then, for all $x, y > 0$,*

$$\mathbb{P}(M_n \geqslant x, \langle M \rangle_n \leqslant y) \leqslant \exp\Big(- \frac{x^2}{2(y + cx)} \Big) .$$

Under some additional assumption, we have the following more useful result:

Theorem 5.10 (*De la Peña's inequality*) *If (M_n) is locally square integrable and conditionally symmetric which means that for $n \geqslant 1$, the conditional distribution of ΔM_n given $\mathcal{F}_{n-1}$ is symmetric, then $\forall x, y > 0$,*

$$\mathbb{P}(M_n \geqslant x, [M]_n \leqslant y) \leqslant \exp\Big(-\frac{x^2}{2y}\Big)$$

Now if $N \in \mathbb{N}$ is a finite stopping time, we want to give a a concentration inequality for the self-normalized martingale $M_N/\sqrt{\langle M\rangle_N}$. However, it is well-known that this is not possible in general (see De la Peña (1999), de la Peña et al. (2004)). Fortunately, the fact that no concentration result applies does not prevent us with getting some useful result. Actually, it is proved in Delattre and Gaiffas (2010) that one can have the following general result (which is not a concentration result):

Theorem 5.11 (*Gaiffas' inequality*) *Let v and x be positive numbers and let N be a finite stopping time. Assume that $M_0 = 0$ and that*

$$\Delta M_n = s_{n-1}\varepsilon_n\,,$$

where $(s_n)_{n\geqslant 0}$ and $(\varepsilon_n)_{n\geqslant 0}$ are $\mathcal{F}_n$-adapted sequences with

$$\mathbb{E}(\varepsilon_n|\mathcal{F}_{n-1}) = 0 \text{ and } |\varepsilon_n| \leqslant b \text{ almost surely for any } n \geqslant 1\,.$$

Then we have the property that

$$\mathbb{P}(|M_N| \geqslant \sqrt{V_N}c_1 b\sqrt{x + c_2\log\Big(\log\big(\frac{v}{V_N}\vee\frac{V_N}{v}\big)\vee e\Big)} \leqslant c(1+e^{b^2})e^{-x}$$

where c, c_1, c_2 are explicit numerical constants and V_n is defined by $V_n \overset{\text{def}}{=} \sum_{i=1}^{n} s_{k-1}^2$.

1.3 Random matrices

We now provide some useful results about random matrices. The first interesting one is given by the following constructive version of the Johnson-Lindenstrauss Lemma (see Dasgupta and Gupta (2003)). Then we detail results about the behavior of singular values of random matrices that are of special interest when solving linear systems.

Johnson-Lindenstrauss Lemma Let A be a $P \times F$ matrix of iid Gaussian $\mathcal{N}(0, 1/P)$ entries. Then the following lemma states that the random (with respect to the choice of the matrix A) variable $||Au||^2$ concentrates around its expectation $||u||^2$ when P is large.

Lemma 5.5 *For any vector u in $\mathbb{R}^F$ and any $\varepsilon \in (0,1)$, we have*

$$\begin{aligned}\mathbb{P}\Big(||Au||^2 \geqslant (1+\varepsilon)||u||^2\Big) &\leqslant e^{-P(\varepsilon^2/4-\varepsilon^3/6)}\,,\\ \mathbb{P}\Big(||Au||^2 \leqslant (1-\varepsilon)||u||^2\Big) &\leqslant e^{-P(\varepsilon^2/4-\varepsilon^3/6)}\,.\end{aligned}$$

The proof directly uses concentration inequalities like Cramér (1938) large deviation Theorem, and may be found e.g. in Achlioptas (2003). Note that the gaussianity is not needed here as only sub-gaussianity is used, and this is also true for other distributions including for instance:

- $\pm$ Rademacher distributions which takes values $\pm 1/\sqrt{P}$ with equal probability $1/2$,
- Distribution taking values $\pm\sqrt{3/P}$ with probability $1/6$ and 0 with probability $2/3$ which produces sparser matrices.

Singular values The asymptotic behavior of the singular values of Random matrices was studied long ago, and very precise a deep results have been found, see Wigner (1958), Marčenko and Pastur (1967), Tao et al. (2010) for instance.

A first useful non-asymptotic property of random matrices is the following result by Edelman (1988) that gives a lower bound of the smallest singular value of A (i.e. the eigenvalue of $\sqrt{A^T A}$), which is of direct interest when considering existence or stability property of least squares estimates.

Theorem 5.12 (*Smallest singular value of Gaussian matrices*) *Let A be an $n \times n$ random matrix whose entries are independent standard normal random variables. Then for every n and fixed $\varepsilon \geqslant 0$ one has the following non-asymptotic bound:*

$$\mathbb{P}(s_{\min}(A) \leqslant \varepsilon n^{-1/2}) \leqslant \varepsilon.$$

Actually one can prove more general results including results for the largest eigenvalue. We refer to Rudelson and Vershynin (2010) for further developments on this topic. The following result holds for arbitrary sub-gaussian matrix, see Rudelson and Vershynin (2008a):

Theorem 5.13 (*Smallest singular value of rectangular random matrices*) *Let A be an $N \times n$ random matrix whose entries are independent and identically distributed subgaussian random variables with zero mean and unit variance. Then for all $\varepsilon \geqslant 0$,*

$$\mathbb{P}(s_{\min}(A) \leqslant \varepsilon(\sqrt{N} - \sqrt{n-1})) \leqslant (C\varepsilon)^{N-n+1} + c^N$$

where $C > 0$ and $c \in (0,1)$ depend only on the subgaussian moment of the entries.

1.4 Talagrand's generic chaining

In order to go beyond the control of a sole random variable, we now quickly describe so-called Talagrand's generic chaining, see Talagrand (2005). Let X be random variable, and $\mathcal{F}$ be a class of functions, centered w.r.t. the law of X. Chaining enables to bound the supremum of the empirical sums over the class $\mathcal{F}$, once we have introduced the distance d such that the following holds true:

$$\mathbb{P}\Big(|\sum_{i=1}^{n}(f_1(X_i) - f_2(X_i))| > t d(f_1, f_2)\Big) \leqslant 2\exp(-t^2/2).$$

Note that for bounded functions, an application of Hoeffding's lemma shows that $d(f_1, f_2)$ is typically of order $\sqrt{n}$. Now, using this distance d, we have the following property:

$$\mathbb{E}\Big(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i)\Big) \quad = \quad O(\frac{1}{n}\int_0^\infty \sqrt{\log(\mathcal{N}(\varepsilon,\mathcal{F},d))}d\varepsilon)$$

where $\mathcal{N}(\varepsilon,\mathcal{F},d)$ is covering number of the class $\mathcal{F}$ with balls of radius ε.

See also Audibert and Bousquet (2007) for a nice survey of related results including results in high probability. For instance let us use the standard notations $P_n f \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ and $Pf = \mathbb{E}(f(X))$. Then we have the following result that holds with probability higher than $1-\delta$:

$$\sup_{f\in\mathcal{F}}\{Pf - P_n f\} \leqslant C\Big(\frac{1}{\sqrt{n}}\mathbb{E}\int_0^\infty \sqrt{\log(\mathcal{N}(\varepsilon,\mathcal{F},d_n))}d\varepsilon + \sqrt{\frac{\log(\delta^{-1})}{n}}\Big)$$

where C is some numerical constant and $d_n(f_1,f_2) \stackrel{\text{def}}{=} \sqrt{P_n(f_1-f_2)^2}$.

There are of course plenty of other results including the use of Rademacher complexities, Bernstein type inequalities, control in other useful norms such as the empirical sup norm, small ball estimates...etc

2 Probably-Approximately-Correct analysis

In this section, we present tools from Probably Approximately Correct (PAC) analysis, which can be seen as an alternative to analysis based on concentration inequalities. This approach has been popularized over the last few years as it leads to tighter results than concentration inequalities based approach as well as simpler proofs. The small disadvantage of these methods is that we need to sample according to a specific (Gibbs) distribution, which may be hard to do in general, but important progress have been made also in this direction; see Narayanan and Rakhlin (2010) for instance when the potential function is convex. Although we do not explicitly make use of these tools in this Ph.D. thesis (but note they are related to the exponentially weighted forecaster presented in chapter 1), we believe they are very important for the development of future Bandit and Reinforcement Learning theory.

2.1 Abc of Pac-analysis

Let us first introduce the following notations, for a given real-valued function f:

$$\mathbb{E}^+ f(X) \stackrel{\text{def}}{=} \log \mathbb{E}\exp(f(X)) \text{ and } \mathbb{E}^- f(X) \stackrel{\text{def}}{=} -\log \mathbb{E}\exp(-f(X))\,.$$

Note that $\mathbb{E}^+ f(X) = -\mathbb{E}^-(-f(X))$. The following immediate property justifies the notation:

Lemma 5.6 (Structural property) *For a given real-valued function f, we have*

$$\mathbb{E}^- f(X) \leqslant \mathbb{E} f(X) \leqslant \mathbb{E}^+ f(X)\,,$$

Moreover, if $\forall x\ 0 \leqslant f(x) \leqslant b$, then we have

$$\mathbb{E}^+ f(X) - \frac{e^b - 1 - b}{b^2}\mathbb{E} f^2(X) \leqslant \mathbb{E} f(X) \leqslant \mathbb{E}^- f(X) + \frac{1}{2}\mathbb{E} f^2(X)\,,$$

where the right hand side inequality holds for non-negative unbounded f as well.

Proof: The proof of the first line is direct by application of Jensen's inequality, the proof of the second line comes from a Taylor expansion of exp, see for instance Lemma 3.3 of Auer et al. (1995) for details. □

The two following inequalities correspond to Lemma 4.2 in Audibert and Catoni (2010b):

Lemma 5.7 *For a given real-valued function of two variables g, we have*

$$\mathbb{E}^+_{x\sim\nu}\mathbb{E}^-_{y\sim\nu'} g(x,y) \leqslant \mathbb{E}^-_{y\sim\nu'}\mathbb{E}^+_{x\sim\nu} g(x,y)\,,$$
$$\mathbb{E}_{x\sim\nu}\mathbb{E}^-_{y\sim\nu'} g(x,y) \leqslant \mathbb{E}^-_{y\sim\nu'}\mathbb{E}_{x\sim\nu} g(x,y)\,.$$

After presenting the properties of $\mathbb{E}^+$ and of $\mathbb{E}^-$, we now present the two key formulas that are used in PAC analysis:

Lemma 5.8 (Key PAC formulas) *Let $\mathcal{C}_b(\mathcal{X})$ be the set of continuous and bounded functions on $\mathcal{X}$. The two formulas hold:*

- (Variational formula) *For any distributions $\nu, \nu' \in \mathcal{P}(\mathcal{X})$, then*

$$\mathcal{K}(\nu,\nu') = \sup\{\mathbb{E}_\nu(f(X)) - \mathbb{E}^+_{\nu'}(f(X)); f \in \mathcal{C}_b(\mathcal{X})\}\,,$$

- (Entropy formula) *For any function $f \in \mathcal{C}_b(\mathcal{X})$,and distribution $\nu' \in \mathcal{P}(\mathcal{X})$, then*

$$\mathbb{E}^+_{\nu'} f(X) = \sup_{\nu\in\mathcal{P}(\mathcal{X})}\{\mathbb{E}_\nu f(X) - \mathcal{K}(\nu,\nu')\}\,.$$

Note that these formulas are generally used with $\mathcal{X}$ being a function space $\mathcal{F}$. Note also that the entropy formula is generally written in the following equivalent form:

$$\exp\{\sup_{\nu\in\mathcal{P}(\mathcal{X})}\{\mathbb{E}_\nu f(X) - \mathcal{K}(\nu,\nu')\}\} = \mathbb{E}_{\nu'}(\exp(f(X))\,.$$

The core of PAC analysis is finally given by the following immediate property:

Lemma 5.9 *Let X be some real-valued random variable. We have the property that whenever $\mathbb{E}^+ X \leqslant 0$, then for all $\delta > 0$, with probability higher than $1-\delta$, it holds that*

$$X \leqslant \log(\delta^{-1})\,.$$

Interestingly, the equivalent of Hoeffding's and Bernstein's inequality can be derived as well for the PAC analysis. The following results are adapted from Alquier (2008):

Lemma 5.10 (PAC Hoeffding's inequality) *For every $a \in (0,1)$, positive function f and n i.i.d. samples $\{X_i\}_{i\leqslant n} \sim \mathcal{P}_X$, we have the property that*

$$\mathbb{E}_X^+\Big(\frac{1}{n}\sum_{i=1}^n \varphi_a(f(X_i)) - \varphi_a(\mathbb{E}_X f(X))\Big) = 0$$

where $\varphi_a(t) = n\log(1 - \frac{t}{n} \wedge a)$.

Lemma 5.11 (PAC Bernstein's inequality) *For every positive function f bounded by b and n i.i.d. samples $\{X_i\}_{i\leqslant n} \sim \mathcal{P}_X$, we have the property that*

$$\mathbb{E}_X^+\Big(\mathbb{E}_X f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{\sigma^2}{n} g(\frac{2b}{n})\Big) \leqslant 0$$

where $\sigma^2 = \mathbb{E}(f^2(X))$ and $g(x) \stackrel{\text{def}}{=} \frac{\exp(x)-1-x}{x^2}$.

2.2 Pac-analysis of regression

In order to motivate the benefit of PAC analysis, we now give a short application to regression, re-deriving one of the results proved in Audibert and Catoni (2010b) for which we hope we provide additional intuition. It is here striking to note the simplicity of the proof as well as the generality and tightness of the bound derived using such technique.

Setting. We consider a regression model of the form $Y = f^\star(X) + \eta$ where $Z = (X, Y)$ follows a low $\mathcal{P}$, $f^\star \in \mathcal{F}$ and η is a centered noise independent from Z. For some positive real-valued loss function l, we introduce $l(f, Z) \in \mathbb{R}^+$ the prediction loss of a function $f \in \mathcal{F}$ for the random variable Z, and then for $f, f' \in \mathcal{F}$, the prediction gap between f and f' $\Delta(f, f') \stackrel{\text{def}}{=} l(f, Z) - l(f', Z)$.

The goal, given an i.i.d $\{Z_i\}_{i\leqslant N}$ sample from $\mathcal{P}$ is to build some probability distribution $\widehat{\pi}$ such that the quantity $\Delta(\widehat{f}, f^\star)$ where $\widehat{f} \sim \widehat{\pi}$ is minimal with high $\mathcal{P} \times \widehat{\pi}$-probability.

Assumptions. We report here the two following assumptions from Audibert and Catoni (2010b):

- (A) There exists some $\eta \in (0,1)$ and a known λ such that for all $f \in \mathcal{F}$,

$$\frac{1}{1+\eta}\mathbb{E}_Z^+\lambda\Delta(f, f^\star) \leqslant \mathbb{E}_Z\lambda\Delta(f, f^\star) \leqslant \frac{1}{1-\eta}\mathbb{E}_Z^-\lambda\Delta(f, f^\star)$$

  This assumption corresponds to a moment assumption (as suggests Lemma 5.6 above); for the square loss $l(f, Z) = (Y - f(X))^2$, Lemma 3.4 of Audibert and Catoni (2010b) shows that it is satisfied whenever $\mathcal{F}$ has finite L^∞-diameter and if $\exists A > 0$ such that the quantity $\mathbb{E}\Big\{\exp\big(A^{-1}|Y - f^\star(X)|\big)|X = x\Big\}$ is uniformly bounded over $\mathcal{X}$.

- (B) There exists a known probability distribution π, and positive constants D, c such that for every $0 < \alpha < \beta$,

$$\mathbb{E}^-_{f\sim\pi}\beta\mathbb{E}_Z\Delta(f,f^\star) - \mathbb{E}^-_{f\sim\pi}\alpha\mathbb{E}_Z\Delta(f,f^\star) \leqslant D\log(\frac{c\beta}{\alpha})\,.$$

  From Lemma 3.3 of Audibert and Catoni (2010b) this assumptions holds for the square loss, and $\mathcal{F}$ being a linear space of dimension d, with π being the uniform distribution on $\mathcal{F}$, $c = 1$ and $D = \frac{d}{2}$.

Algorithm. The main idea is the use of a Gibbs distribution. For a reference measure π on $\mathcal{F}$, and some function $\xi : \mathcal{F} \to \mathbb{R}$, we define $\pi_{-\xi}$ to be the distribution with density w.r.t. π:

$$\frac{d\pi_{-\xi}}{d\pi}(f) = \frac{\exp(-\xi(f))}{\int \exp(-\xi(f'))\pi(df')}\,.$$

Note that this naturally involves $\mathbb{E}^-$ as can be seen from the relation

$$\log(\frac{d\pi_{-\xi}}{d\pi}(f)) = \mathbb{E}^-_\pi\xi(f) - \xi(f)\,.$$

Note also that since $\mathbb{E}^+_{\pi_{-\xi}}\xi(f) = \log(\pi(\mathcal{F})) + \mathbb{E}^-_\pi(\xi(f))$, we naturally have for all $\delta \in (0,1)$:

$$\mathbb{P}_{\pi_{-\xi}}\Big(\xi(f)) \leqslant \log(\pi(\mathcal{F})) + \mathbb{E}^-_\pi\xi(f) + \log(1/\delta)\Big) \geqslant 1-\delta\,.$$

We consider from now on that the $\widehat{\pi}$ is the specific Gibbs distribution defined using $\xi(f) = -n\lambda P_n l(f,Z)$ where λ is the constant of hypothesis (A), which is also the same as with $\xi(f) \stackrel{\text{def}}{=} -n\lambda P_n\Delta(f,f^\star)$. We introduce for symmetry reasons the Gibbs distribution π_0 defined using $\xi(f) \stackrel{\text{def}}{=} -n\lambda_0\mathbb{E}_Z\Delta(f,f^\star)$ for some constant λ_0 that has to be tuned.

Theorem 5.14 (*PAC risk bounds for regression*) *Under the two previous assumptions, the random estimate $\widehat{f} \sim \widehat{\pi}$ satisfies for all $0 < \lambda_0 < (1-\eta)\lambda$ with $\mathcal{P}\times\widehat{\pi}$-probability higher than $1-\delta$*

$$\mathbb{E}_Z\Delta(\widehat{f},f^\star) \leqslant \frac{1}{n[(1-\eta)\lambda - \lambda_0]}\Big[D\log\Big(\frac{c(1+\eta)\lambda}{\lambda_0}\Big) + 2\log(2/\delta)\Big]\,.$$

Proof: **Step 1. Empirical and expected gap.** We have the property that for a fixed f, $\mathbb{E}^-_Z(nP_n\Delta(f,f^\star)) = n\mathbb{E}^-_Z\Delta(f,f^\star)$, which successively entails that

$$\begin{aligned}
&\mathbb{E}^-_Z\big[n\lambda P_n\Delta(f,f^\star) - n\mathbb{E}^-_Z\lambda\Delta(f,f^\star)\big] = 0\,,\\
\text{then}\quad &\mathbb{E}^+_Z\big[n\mathbb{E}^-_Z\lambda\Delta(f,f^\star) - n\lambda P_n\Delta(f,f^\star)\big] = 0\,,\\
\text{then}\quad &\mathbb{E}^+_{(Z,f)\sim\mathcal{P}\times\pi_0}\big[n\mathbb{E}^-_Z\lambda\Delta(f,f^\star) - n\lambda P_n\Delta(f,f^\star)\big] = 0\,,
\end{aligned}$$

where we used in the second line that $\mathbb{E}^+_Z(-f) = -\mathbb{E}^-(f)$, and in the third line that π_0 is a distribution that does not dependent on the random variable Z.

Step 2. The benefit of using a Gibbs measure. Now we can not use this formula for our distribution $\widehat{\pi}$ since it depends on the sample explicitly. In order to overcome this, we remark that when $\pi_0 << \widehat{\pi}$, then the previous formula can be rewritten

$$\mathbb{E}^+_{(Z,\widehat{f})\sim\mathcal{P}\times\widehat{\pi}}\big[n\mathbb{E}^-_Z\lambda\Delta(\widehat{f},f^\star)-n\lambda P_n\Delta(\widehat{f},f^\star)+\log\frac{d\pi_0}{d\widehat{\pi}}(\widehat{f})\big]=0\,.$$

and in the general case the equality becomes an inequality.

Thus this simple remark together with the definition of the Gibbs measure prove that $\mathbb{E}^+_{(Z,\widehat{f})\sim\mathcal{P}\times\widehat{\pi}}\big(V(Z,\widehat{f})\big)\leqslant 0$, i.e. that $V(Z,\widehat{f})\leqslant\log(1/\delta)$ with $\mathcal{P}\times\widehat{\pi}$-probability higher than $1-\delta$, where we introduced for convenience the quantity

$$\begin{aligned}V(Z,\widehat{f}) \;\stackrel{\text{def}}{=}\; & [n\mathbb{E}^-_Z\lambda\Delta(\widehat{f},f^\star)-n\lambda P_n\Delta(\widehat{f},f^\star)]+[nP_n\lambda\Delta(\widehat{f},f^\star)-\mathbb{E}^-_\pi n\lambda P_n\Delta(f,f^\star)]\\ & +[\mathbb{E}^-_\pi n\lambda_0\mathbb{E}_Z\Delta(f,f^\star)-n\lambda_0\mathbb{E}_Z\Delta(\widehat{f},f^\star)]\,,\end{aligned}$$

which further simplifies into

$$V(Z,\widehat{f})=[n\mathbb{E}^-_Z\lambda\Delta(\widehat{f},f^\star)-\mathbb{E}^-_\pi n\lambda P_n\Delta(f,f^\star)]+[\mathbb{E}^-_\pi n\lambda_0\mathbb{E}_Z\Delta(f,f^\star)-n\lambda_0\mathbb{E}_Z\Delta(\widehat{f},f^\star)]\,.$$

Step 3. Cleaning the bound. Now we can further remove the last occurrence of P_n in $V(Z,\widehat{f})$ by noticing that since $\mathbb{E}^+_Z\mathbb{E}^-_\pi n\lambda P_n\Delta(f,f^\star)\leqslant\mathbb{E}^-_\pi\mathbb{E}^+_Z n\lambda P_n\Delta(f,f^\star)$, and since the two terms of this inequality do not depend on $\widehat{f}$, then we further have $\mathbb{E}^+_{Z,\widehat{f}}[\mathbb{E}^-_\pi n\lambda P_n\Delta(f,f^\star)]\leqslant\mathbb{E}^-_\pi n\mathbb{E}^+_Z\lambda\Delta(f,f^\star)$. Thus by application of Lemma 5.9, we deduce that with $\mathcal{P}\times\widehat{\pi}$-probability higher than $1-\delta$, then $V_2(Z,\widehat{f})\leqslant\log(1/\delta)$, where we introduced

$$V_2(Z,\widehat{f})\stackrel{\text{def}}{=}\mathbb{E}^-_\pi n\lambda P_n\Delta(f,f^\star)-\mathbb{E}^-_\pi n\mathbb{E}^+_Z\lambda\Delta(f,f^\star)\,.$$

Thus, combining the bound on $V(Z,\widehat{f})$ together with the bound on $V_2(Z,\widehat{f})$ with a union bound, so far we have proved that with $\mathcal{P}\times\widehat{\pi}$-probability higher than $1-\delta$,

$$n\mathbb{E}^-_Z\lambda\Delta(\widehat{f},f^\star)-n\lambda_0\mathbb{E}_Z\Delta(\widehat{f},f^\star)\leqslant\mathbb{E}^-_\pi n\mathbb{E}^+_Z\lambda\Delta(f,f^\star)-\mathbb{E}^-_\pi n\lambda_0\mathbb{E}_Z\Delta(f,f^\star)+2\log(2/\delta)\,.$$

Step 4. Applying assumptions (A) and (B). Now, by applying assumption (A), we conclude that with $\mathcal{P}\times\widehat{\pi}$-probability higher than $1-\delta$,

$$[n(1-\eta)\lambda-n\lambda_0]\mathbb{E}_Z\Delta(\widehat{f},f^\star)\leqslant\mathbb{E}^-_\pi n(1+\eta)\lambda\mathbb{E}_Z\Delta(f,f^\star)-\mathbb{E}^-_\pi n\lambda_0\mathbb{E}_Z\Delta(f,f^\star)+2\log(2/\delta)\,,$$

and then by applying assumption (B), we deduce the following final result that holds for all $0<\lambda_0<(1-\eta)\lambda$:

$$n[(1-\eta)\lambda-\lambda_0]\mathbb{E}_Z\Delta(\widehat{f},f^\star)\leqslant D\log\Big(\frac{c(1+\eta)\lambda}{\lambda_0}\Big)+2\log(2/\delta)\,.$$

□

Of course, PAC analysis can be combined with tools like generic chaining (see Audibert and Bousquet (2007)), and can also be used for model selection and more generally for what is called model aggregation (see Dalalyan and Tsybakov (2008)). Moreover, as suggests the use of exponentially weighted forecasters for bandits, see chapter 1, there are indeed strong links between PAC analysis and bandit theory. Such links begin to be studied, see Seldin et al. (2011a,b) for important introductory work on this subject and will be developed in a near future.

Chapter 6

Linear Regression with Random Projections.

We investigate a method for regression that makes use of a randomly generated subspace $\mathcal{G}_P \subset \mathcal{F}$ (of finite dimension P) of a given large (possibly infinite) dimensional function space $\mathcal{F}$, e.g. $L_2([0,1]^d; \mathbb{R})$. $\mathcal{G}_P$ is defined as the span of P random features that are linear combinations of a basis functions of $\mathcal{F}$ weighted by random Gaussian i.i.d. coefficients. We show practical motivation for the use of this approach, detail the link that this random projections method shares with RKHS and Gaussian objects theory and prove, both in deterministic and random design, approximation error bounds when searching for the best regression function in $\mathcal{G}_P$ rather than in $\mathcal{F}$ and derive excess risk bounds for a specific regression algorithm (least squares regression in $\mathcal{G}_P$). This chapter stresses the motivation to study such methods, thus the analysis developed is kept simple for explanations purpose and leaves room for future developments.

The work presented here corresponds to two articles published in the proceedings of the 23*rd and* 24*th conferences on advances in Neural Information Processing Systems (NIPS)*, see Maillard and Munos (2009), Maillard and Munos (2010a) and one under review in the *Journal of Machine Learning Research (JMLR)*. I would like to thank here *Pierre Chainais* and *Olivier Degris* for interesting pointers to the literature in image processing and applied reinforcement learning.

Contents

1 Introduction

We consider a standard regression problem. Thus let us introduce $\mathcal{X}$ an input space, and $\mathcal{Y} = \mathbb{R}$ the real line. We denote by $\mathcal{P}$ an unknown probability distribution over the product space $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and by $\mathcal{P}_{\mathcal{X}}$ its first marginal, i.e. $d\mathcal{P}_{\mathcal{X}}(x) = \int_{\mathbb{R}} \mathcal{P}(x, dy)$. In order for this quantity to be well defined we assume that $\mathcal{X}$ is a Polish space (i.e., metric, complete, separable), see (Dudley, 1989, Th. 10.2.2.). Finally, let $L_{2,\mathcal{P}_{\mathcal{X}}}(\mathcal{X}; \mathbb{R})$ be the space of real-valued functions on $\mathcal{X}$ that are squared integrable with respect to (w.r.t.) $\mathcal{P}_{\mathcal{X}}$, equipped with the quadratic norm

$$\|f\|_{\mathcal{P}_{\mathcal{X}}} \stackrel{\text{def}}{=} \sqrt{\int_{\mathcal{X}} f^2(x) dP_{\mathcal{X}}(x)}\,.$$

In this chapter, we consider that $\mathcal{P}$ has some structure corresponding to a model of regression with random design; there exists a (unknown) function $f^\star : \mathcal{X} \to \mathbb{R}$ such that if $(x_n, y_n)_{n \leqslant N} \in \mathcal{X} \times \mathbb{R}$ are independently and identically distributed (i.i.d.) according to $\mathcal{P}$, then one can write

$$y_n = f^\star(x_n) + \eta_n\,,$$

where η_n is a centered noise, independent from $\mathcal{P}_{\mathcal{X}}$, introduced for notational convenience. In terms of random variables, we will often simply write $Y = f^\star(X) + \eta$ where $(X, Y) \sim \mathcal{P}$.

Let $\mathcal{F} \subset L_{2,\mathcal{P}_{\mathcal{X}}}(\mathcal{X}; \mathbb{R})$ be some given class of functions. The goal of the statistician is to build, from the observations only, a regression function $\widehat{f} \in \mathcal{F}$ that is closed to the so-called target function $f^\star$, in the sense that it has a low excess risk $R(f) - R(f^\star)$, where the risk of any $f \in L_{2,\mathcal{P}_{\mathcal{X}}}(\mathcal{X}; \mathbb{R})$ is defined as

$$R(f) \stackrel{\text{def}}{=} \int_{\mathcal{X}\times\mathbb{R}} (y - f(x))^2 d\mathcal{P}(x, y)$$

Similarly, we introduce the empirical risk of a function f to be

$$R_N(f) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} [y_n - f(x_n)]^2 .$$

and we define the empirical norm of f as $\| f \|_N \stackrel{\text{def}}{=} \sqrt{\frac{1}{N} \sum_{n=1}^{N} f(x_n)^2}$.

Function spaces and penalization. In this work, we consider that $\mathcal{F}$ is an infinite dimensional space that is generated by the span over a denumerable family of functions $\{\varphi_i\}_{i\geqslant 1}$ of $L_{2,\mathcal{P}_{\mathcal{X}}}(\mathcal{X}; \mathbb{R})$: We call the $\{\varphi_i\}_{i\geqslant 1}$ the *initial features* and thus refer to $\mathcal{F}$ as to the initial feature space:

$$\mathcal{F} \stackrel{\text{def}}{=} \{f_\alpha(x) \stackrel{\text{def}}{=} \sum_{i\geqslant 1} \alpha_i \varphi_i(x), \|\alpha\| < \infty\}.$$

Examples of initial features include Fourier basis, multi-resolution basis such as wavelets, and also less explicit features coming from a preliminary dictionary learning process.

In the sequel, for the sake of simplicity we focus our attention to the case when the target function $f^\star = f_{\alpha^\star}$ belongs to the space $\mathcal{F}$, in which case the excess risk of a function f can be written as $R(f) - R(f^\star) = \|f - f^\star\|_{\mathcal{P}_{\mathcal{X}}}$. Since $\mathcal{F}$ is an infinite dimensional space, empirical risk minimization in $\mathcal{F}$ defined by $\arg\min_{f\in\mathcal{F}} R_N(f)$ is certainly subject to overfitting. Traditional methods to circumvent this problem consider penalization techniques, i.e. one searches for a function that satisfies

$$\widehat{f} = \arg\min_{f\in\mathcal{F}} R_N(f) + \text{pen}(f),$$

where typical examples of penalization include $\text{pen}(f) = \lambda \|f\|_p^p$ for $p = 1$ or 2, where λ is a parameter and usual choices for the norm are ℓ_2 (ridge-regression Tikhonov (1963)) and ℓ_1 (LASSO Tibshirani (1994)).

Motivation. In this chapter we follow a complementary approach introduced in Maillard and Munos (2009) for finite dimensional space, called Compressed Least Squares Regression, and extended in Maillard and Munos (2010a), which considers generating *randomly* a space $\mathcal{G}_P \in \mathcal{F}$ of finite dimension P and then returning an empirical estimate in $\mathcal{G}_P$. The empirical risk minimizer in $\mathcal{G}_P$, i.e. $\arg\min_{g\in\mathcal{G}_P} R_N(g)$ is a natural candidate, but other choices of

estimates are possible, based on traditional literature on regression when $P < N$ (penalization, projection, PAC-Bayesian estimates...) The generation of the space $\mathcal{G}_P$ makes use of random matrices, that have already demonstrated their benefit in different settings (see for instance Zhao and Zhang (2009) about spectral clustering or Dasgupta and Freund (2008) about manifold learning).

Our goal is first to give some intuition about this method by providing approximation error and simple excess risk bounds (which may not be the tightest possible ones as explained in Section 4.2) for the proposed method, and also by providing links to other standards approaches, in order to encourage research in that direction, which, as showed in the next section, has already been used in several applications.

Outline of the chapter.

In Section 2, we quickly present the method and give practical motivation for investigating this approach. In Section 3, we give a short overview of Gaussian objects theory (subsection 3.1), which enables us to show how to relate the choice of the initial features $\{\varphi_i\}_{i\geqslant 1}$ to the construction of standard function spaces via Gaussian objects (subsection 3.2), and we finally state a useful version of the Johnson-Lindenstrauss Lemma for our setting (subsection 3.3).

In Section 4, we describe a typical algorithm (subsection 4.1), and then provide some quick survey of classical results in regression while discussing the validity of their assumptions in our setting (subsection 4.2). Then our main results are stated in subsection 4.3, where we provide bounds on approximation error of the random space $\mathcal{G}_P$ in the framework of regression with deterministic and random designs, and in subsection 4.4, where we derive excess risk bounds for some specific estimate.

Section 5 provide some discussion about existing results and finally appendix 6 contains the proofs.

2 Summary of the method

From now on, we assume that the set of features $\{\varphi_i\}_{i\geqslant 1}$ are continuous and satisfy the assumption that,

$$\sup_{x\in\mathcal{X}} \|\varphi(x)\|^2 < \infty, \text{ where } \varphi(x) \stackrel{\text{def}}{=} (\varphi_i(x))_{i\geqslant 1} \in l_2 \text{ and } \|\varphi(x)\|^2 \stackrel{\text{def}}{=} \sum_{i\geqslant 1} \varphi_i(x)^2.$$

Let us introduce a set of P *random features* $(\psi_p)_{1\leqslant p\leqslant P}$ defined as linear combinations of the initial features $\{\varphi_i\}_{1\geqslant 1}$ weighted by random coefficients:

$$\psi_p(x) \stackrel{\text{def}}{=} \sum_{i\geqslant 1} A_{p,i}\varphi_i(x), \text{ for } 1 \leqslant p \leqslant P \tag{6.1}$$

where the (infinitely many) coefficients $A_{p,i}$ are drawn i.i.d. from a centered distribution (e.g. Gaussian) with variance $1/P$. Then let us define $\mathcal{G}_P$ to be the (random) vector space

spanned by those features, i.e.

$$\mathcal{G}_P \stackrel{\text{def}}{=} \{g_\beta(x) \stackrel{\text{def}}{=} \sum_{p=1}^{P} \beta_p \psi_p(x), \beta \in \mathbb{R}^P\}.$$

From now on, $\mathcal{P}_\mathcal{G}$ will refer to the law of the Gaussian variables, $\mathcal{P}_\eta$ to the law of the observation noise and $\mathcal{P}_\mathcal{Y}$ to the law of the observations. Remember also that $\mathcal{P}_\mathcal{X}$ refers to the law of the inputs.

One may naturally wish to build an estimate $g_{\widehat{\beta}}$ in the linear space $\mathcal{G}_P$. For instance in the case of deterministic design, if we consider the ordinary least squares estimate, i.e. $\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^P} R_N(g_\beta)$, then we can derive the following result (see Section 4.4 for a similar result with random design):

Theorem 6.1 (Deterministic design) *Assuming that the random variable Y is such that $|Y| \leqslant B$, then for all $P \geqslant 1$, for all $\delta \in (0,1)$ there exists an event of $\mathcal{P}_\mathcal{Y} \times \mathcal{P}_\mathcal{G}$-probability larger than $1-\delta$ such that on this event, the excess risk of the least squares estimate $g_{\widehat{\beta}}$ is bounded as*

$$\|g_{\widehat{\beta}} - f^\star\|_N^2 \leqslant \frac{12\log(8N/\delta)}{P}\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^{N}\|\varphi(x_n)\|^2 + \kappa B^2 \frac{P+\log(2/\delta)}{N} \tag{6.2}$$

for some numerical constant $\kappa > 0$.

Example: Let us consider as an example the features $\{\varphi_i\}_{i\geqslant 1}$ to be a set of functions defined by rescaling and translation of a mother one-dimensional hat function (illustrated in Figure 6.1, middle column) and defined precisely in paragraph 3.2.2. Then in this case we can show that

$$\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^{N}\|\varphi(x_n)\|^2 \leqslant \frac{1}{2}\|f^\star\|_{H^1}^2 ,$$

where $H^1 = H^1([0,1])$ is the Sobolev space of order 1. Thus we deduce that the excess risk is bounded as $\|g_{\widehat{\beta}} - f^\star\|_N^2 = O(\frac{B\|f^\star\|_{H^1}\log(N/\delta)}{\sqrt{N}})$ for P of the order $\sqrt{N}$.

Similarly, the analysis given in paragraph 3.2.1 below shows that when the features $\{\varphi_i\}_{i\geqslant 1}$ are wavelets rescaled by a factor $\sigma_i = \sigma_{j,l} = 2^{-js}$ for some real number $s > 1/2$, where j, l are the scale and position index corresponding to the ith element of the family, and that the mother wavelet enables to generate the Besov space $\mathcal{B}_{s,2,2}([0,1])$ (see paragraph 3.2.1), then for some constant c,

$$\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^{N}\|\varphi(x_n)\|^2 \leqslant \frac{c}{1-2^{-2s+1}}\|f^\star\|_{s,2,2}^2 ,$$

Thus the excess risk in this case is bounded as $\|g_{\widehat{\beta}} - f^\star\|_N^2 = O(\frac{B\|f^\star\|_{s,2,2}\log(N/\delta)}{\sqrt{N}})$.

2.1 Comments

The second term in the bound (6.2) is a usual estimation error term in regression, while the first term comes from the additional approximation error of the space $\mathcal{G}_P$ w.r.t. $\mathcal{F}$. It involves the norm of the parameter $\alpha^\star$, and also the norm $\|\varphi(x)\|$ at the sample points.

The nice aspects of this result:

- The weak dependency of this bound with the dimension of the initial space $\mathcal{F}$. This appears implicitly in the terms $\|\alpha^\star\|^2$ and $\frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2$, and we will show that for a large class of function spaces, these terms can be bounded by a function of the norm of $f^\star$ only.

- The result does not require any specific smoothness assumptions on the initial features $\{\varphi_i\}_{i\geqslant 1}$; by optimizing over P, we get a rate of order $N^{-1/2}$ that corresponds to the *minimax* rates under such assumptions up to logarithmic factors.

- Because the choice of the subspace $\mathcal{G}_P$ within which we perform the least-squares estimate is random, we avoid (with high probability) degenerated situations where the target function $f^\star$ cannot be well approximated with functions in $\mathcal{G}_P$. Indeed, in methods that consider a given (deterministic) finite-dimensional space $\mathcal{G}$ of the big space $\mathcal{F}$ (like linear approximation using a predefined set of wavelets), it is often possible to find a target function $f^\star$ such that $\inf_{g\in\mathcal{G}_P}\|f^\star - g\|_N$ is large. Whereas using this method, the random choice of $\mathcal{G}_P$ implies that for any $f^\star \in \mathcal{F}$, the approximation error $\inf_{g\in\mathcal{G}_P}\|f^\star - g\|_N$ can be controlled (by the first term of the bound (6.2)) in high probability. See section 5.2 for an illustration of this property. Thus the results we obtain compete with non-linear approximation (Barron et al., 2008) or kernel ridge regression (Caponnetto and De Vito, 2007).

- In terms of numerical complexity, this approach is more efficient than non-linear regression and kernel ridge regression. Indeed, once the random space has been generated, we simply solve a least squares estimate in a low-dimensional space. The computation of the Gram matrix involves performing random projections (which can be computed efficiently for several choices of the random coefficients $A_{p,i}$, see Liberty et al. (2008), Ailon and Chazelle (2006), Sarlos (2006) and many other references therein). Numerical aspects of the algorithms are described in Section 5.4.

Possible improvements: As mentioned previously we do not make specific assumptions about the initial features $\{\varphi_i\}_{i\geqslant 1}$. However, considering smoothness assumptions on the features would enable to derive a better approximation error term (first term of the bound (6.2)); typically with a Sobolev assumption or order s, we would get a term of order P^{-2s} instead of P^{-1}. For simplicity of the presentation, we do not consider such assumptions here and report the general results.

The $\log(N)$ factor may be seen as unwanted and one would like to remove it. However, this term comes from a variant of the Johnson-Lindenstrauss lemma combined with a union

bound, and it seems difficult to remove it, unless the dimension of $\mathcal{F}$ is small (we can then use covers) which is non interesting for our purpose.

Possible extensions of the random projection method. It seems natural to consider other constructions than the use of i.i.d. Gaussian random coefficients. For instance we may consider Gaussian variables with variance σ_i^2/P different for each i instead of homeoscedastic variables, which is actually equivalent to considering the features $\{\varphi'_i\}_{i\geqslant 1}$ with $\varphi'_i = \sigma_i \varphi_i$ instead.

Although in this work we develop results using Gaussian random variables, such method will essentially work similarly for matrices with sub-Gaussian entries as well.

A more important modification of the method would be to consider, like for data-driven penalization techniques, a data-dependent construction of the random space $\mathcal{G}_P$, i.e. using a data-driven distribution for the random variable $A_{p,i}$ instead of a Gaussian distribution. However such modification *will not* work for the method developed in this chapter and would require a different analysis.

Illustration. In order to illustrate the method, we show in figure 6.1 three examples of initial features $\{\varphi_i\}$ (top row) and random features $\{\psi_p\}$ (bottom row). The first family of features is the basis of wavelet Haar functions. The second one consists of multi-resolution hat functions (see paragraph 3.2.2) and the last one shows multi-resolution Gaussian functions. For example, in the case of multi-resolution hat functions (middle column), the corresponding (random features) are Brownian motions. The linear regression with random projections approach described here simply consists in performing least-squares regression using the set of randomly generated features $\{\psi_p\}_{1\leqslant p\leqslant P}$ (e.g. Brownian motions).

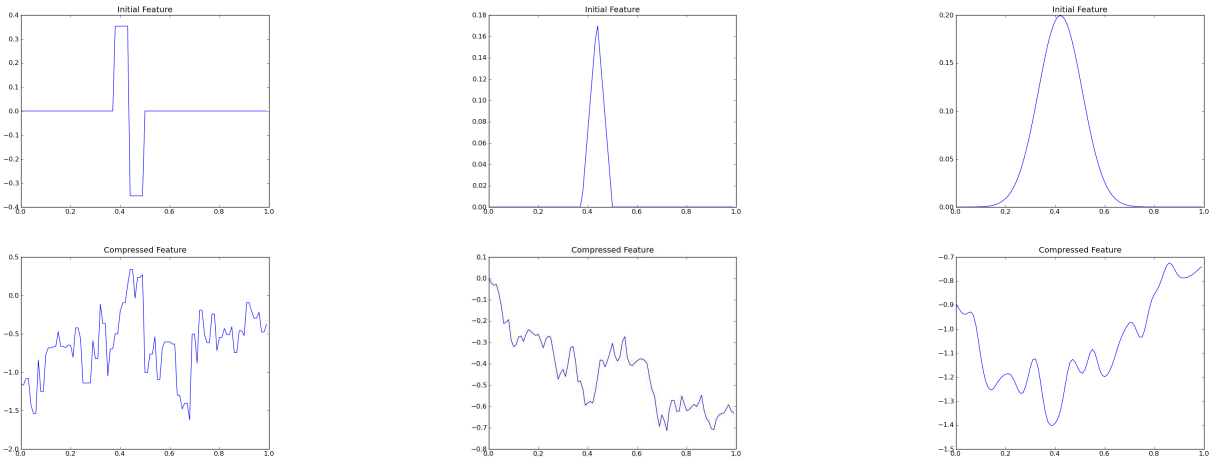


Figure 6.1: Three representative of initial features φ (top row) and a sample of a corresponding random feature ψ (bottom row). The initial set of features are (respectively) Haar functions (left), multi-resolution hat functions (middle) and multi-resolution Gaussian functions (right).

2.2 Motivation from practice

We conclude this introduction with some additional motivation to study such objects coming from practical applications. Let us remind that the use of random projections is well-known in many domains and applications, with different names according to the corresponding field, and the corresponding random objects are widely studied and used. Our contribution is their analysis in a regression setting.

For instance, in Sutton and Whitehead (1993) the authors mentioned such constructions under the name random representation as a tool for value function approximation in practical implementations of reinforcement learning algorithms, and demonstrated the benefit of such methods. They also pointed out that such representations were already used in 1962 in Rosenblatt's perceptron as a preprocessing layer. See also Sutton (1996) for other comments concerning the practical benefit of "random collapsing" methods.

Another example in image processing, when the initial features are chosen to be a wavelet (rescaled) system, the corresponding random features $\{\psi_p\}_{1\leqslant p\leqslant P}$ are special cases of Random Wavelet Series, that are well studied in signal processing and mathematical physics (see Aubry and Jaffard (2002), Durand (2008) for a study of the law of the spectrum of singularities of these series).

Noise model and texture generation: The construction of Gaussian objects (see paragraph 3.2.1) is highly flexible and enables to do automatic noise-texture generation easily, as explained in Deguy and Benassi (2001). In their paper, the authors show that with the appropriate choice of the wavelet functions and when using rescaling coefficients of the form $\sigma_{j,l} = 2^{-js}$ with scale index j an position index l (see paragraph 3.2.1), where s is not a constant but is now a function of j and l, we can generate fractional Brownian motions, multi-scale fractional Brownian motions, and more generally what is called intermittent locally self-similar Gaussian processes.

In particular, for image texture generation they introduce a class of functions called *morphlets* that enables to perform approximations of intermittent locally self-similar Gaussian processes. These approximations are both numerically very efficient and have visually imperceptible differences to the targetted images. The authors also allow other distributions than the Gaussian for the random variables ξ (which thus does not fit the theory presented here), and use this additional flexibility to produce an impressive texture generator.

Figure 6.2 illustrates an example performed on some simple texture model[1] where an image of size 512×512 is generated (two-dimensional Brownian sheet with Hurst index $H = 1.1$) (left) and then subsampled at 32×32 (middle), which provides the data samples for generating a regression function (right) using random features (generated from the symlets as initial features, in the simplest model when s is constant).

[1]The authors wish to thank Pierre Chainais for performing experimental study of random projection methods applied to image processing, and for providing us with interesting pointers to related works.

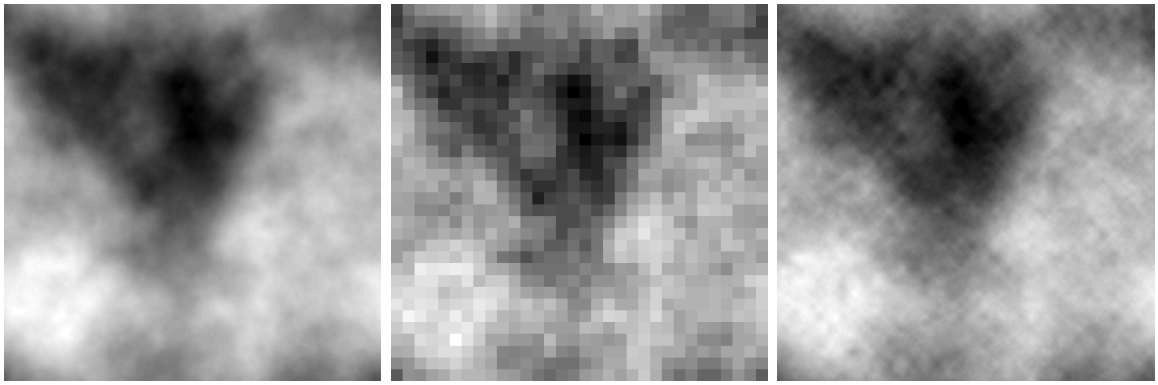


Figure 6.2: Example of an initial large texture (left), subsampled (middle), and possible recovery using regression with random projections (right)

3 Gaussian objects

We now describe some tools of Gaussian object theory that would be useful in later analysis of the method. Each random feature ψ_p built from equation (6.1), when the coefficients are Gaussian, qualifies as a Gaussian object. It is thus natural to study some important features of Gaussian objects.

3.1 Reminder of Gaussian objects theory

In all this section, $\mathcal{S}$ will refer to a vector space, $\mathcal{S}'$ to its topological dual, and $(\cdot, \cdot)$ to its duality product.

Definition 6.1 (Gaussian objects) *A random variable $W \in \mathcal{S}$ is called a Gaussian object if for all $\nu \in \mathcal{S}'$, (ν, W) is a Gaussian (real-valued) variable. We further call any $a \in \mathcal{S}$ to be an* expectation *of W if*

$$\forall \nu \in \mathcal{S}' \, , \, \mathbb{E}(\nu, W) = (\nu, a) < \infty \, ,$$

and any $K : \mathcal{S}' \to \mathcal{S}$ to be a covariance operator *of W if*

$$\forall \nu, \nu' \in \mathcal{S}' \, , \, Cov((\nu, W)(\nu', W)) = (\nu, K\nu') < \infty \, ,$$

where Cov refer to the correlation between two real-valued random variables.

Whenever there exist such a and K, we say that W follows the law $\mathcal{N}(a, K)$. Moreover, W is called a centered Gaussian object if $a = 0$.

Kernel space. We only provide a brief introduction to this notion and refer the interested reader to Lifshits (1995) or Janson (1997) for refinements.

Let $I' : \mathcal{S}' \to L^2(\mathcal{S}, \mathcal{N}(0,K))$ be the canonical injection from the space of continuous linear functionals $\mathcal{S}'$ to the space of measurable linear functionals

$$L_2(\mathcal{S};\mathbb{R},\mathcal{N}(0,K)) = \left\{ z : \mathcal{S} \to \mathbb{R}, \mathbb{E}_{W\sim\mathcal{N}(0,K)}|z(W)|^2 < \infty \right\}.$$

endowed with inner product $\langle z_1, z_2 \rangle = \mathbb{E}(z_1(W)z_2(W))$. For any $\nu \in \mathcal{S}'$, it is defined by $I'(\nu) = (\nu, \cdot)$, which belongs to $L_2(\mathcal{S};\mathbb{R},\mathcal{N}(0,K))$ since by definition of K we have $(\nu, K\nu) = \mathbb{E}(\nu, W)^2 < \infty$.

Then note that the space defined by $\mathcal{S}'_{\mathcal{N}} \stackrel{\text{def}}{=} \overline{I'(\mathcal{S}')}$, i.e. the closure of the image of $\mathcal{S}'$ by I' in the sense of $L_2(\mathcal{S};\mathbb{R},\mathcal{N}(0,K))$, is a Hilbert space with inner product inherited from $L_2(\mathcal{S};\mathbb{R},\mathcal{N}(0,K))$.

Now under the assumption that I' is continuous (see Section 4.1 for practical conditions ensuring that this is the case), then we define the adjoint $I : \mathcal{S}'_{\mathcal{N}} \to \mathcal{S}$ of I', by duality. Indeed for any $\mu \in \mathcal{S}'$ and $z \in I'(\mathcal{S}')$, we have by definition that

$$(\mu, Iz) = \langle I'\mu, z \rangle_{\mathcal{S}'_{\mathcal{N}}} = \mathbb{E}_W((\mu, W)z(W))$$

from which we deduce by continuity that $Iz = \mathbb{E}_W(Wz(W))$. For the sake of clarity, this specifies for instance in the case when $\mathcal{S} = L_2(\mathcal{X};\mathbb{R})$, for all $x \in \mathcal{X}$ as

$$(Iz)(x) = \mathbb{E}_W(W(x)z(W)).$$

Definition 6.2 (Kernel space) *Provided that the mapping I' is continuous, then we define the* **kernel space** *of a centered Gaussian object W as $\mathcal{K} \stackrel{\text{def}}{=} I(\overline{I'(\mathcal{S}')}) \subset \mathcal{S}$.*

The kernel space can be built alternatively based on a separable Hilbert space $\mathcal{H}$ as follows (Lifshits, 1995):

Lemma 6.1 (Construction of the Kernel space.) *Let $J : \mathcal{H} \to \mathcal{S}$ be an injective linear mapping such that $K = JJ'$, where J' is the adjoint operator of J. Then the kernel space of $\mathcal{N}(0,K)$ is $\mathcal{K} = J(\mathcal{H})$, endowed with inner product $\langle Jh_1, Jh_2 \rangle_H \stackrel{\text{def}}{=} \langle h_1, h_2 \rangle_{\mathcal{H}}$.*

We conclude this section with the following Lemma from Lifshits (1995) that enables to define the expansion of a Gaussian object W.

Lemma 6.2 (Expansion of a Gaussian object) *Let $(\varphi_i)_i$ be an orthonormal basis of $\mathcal{K}$ for the inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ and $\{\xi_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0,1)\}_{i\geqslant 1}$. Then $\sum_{i=1}^{\infty} \xi_i \varphi_i$ is a Gaussian object following the law $\mathcal{N}(0,K)$. It is called an* **expansion** *for $\mathcal{N}(0,K)$.*

Note that from Lemma 6.1, one can build an orthonormal basis $(\varphi_i)_i$ by defining, for all $i \geqslant 1$, $\varphi_i = Jh_i$ where $(h_i)_i$ is an orthonormal basis of $\mathcal{H}$ and J satisfies conditions of Lemma 6.1.

3.2 Interpretation of some function spaces with Gaussian objects theory

In this section, we precise the link between Gaussian objects theory and reproducing kernel Hilbert spaces (RKHS) in order to provide more intuition about such objects. Indeed in many cases, the kernel space of a Gaussian object is RKHS. Note, however, that in general, depending on the Gaussian object considered, the former space may also be a more general space (see (Canu et al., 2009) about RKS) when the Hilbert assumption is dropped, therefore, there is no one-to-one correspondence between RKHS and kernel spaces of Gaussian objects and it is worth explaining when the two notions coincide. More importantly, this section shows various examples of classical function spaces, related to the construction of the space $\mathcal{G}_P$ for different choices of initial features $\{\varphi_i\}_{i\geqslant 1}$, which can be useful for applications.

3.2.1 Gaussian objects with a supporting Hilbert space

In this subsection only, we make the assumption that $\mathcal{S} = \mathcal{H}$ is a Hilbert space and we introduce $\{e_i\}_{i\geqslant 1}$ an orthonormal basis of $\mathcal{H}$. Let us now consider $\xi_i \sim \mathcal{N}(0,1)$ i.i.d., and positive coefficients $\sigma_i \geqslant 0$ such that $\sum_i \sigma_i^2 < \infty$. Since $\sum_i \sigma_i^2 < \infty$, the Gaussian object $W = \sum_i \xi_i \sigma_i e_i$ is well defined and our goal is to identify the kernel of the law of W.

To this aim we identify the functions I' and I. Since $\mathcal{S}$ is a Hilbert space, then $\mathcal{S}' = \mathcal{S}$, thus we consider $f = \sum_i c_i e_i \in \mathcal{S}'$ for some $c \in l_2$. For such an f, we deduce that the injection mapping is given by $(I'f)(g) = \sum_i c_i (g, e_i)$, and that we have

$$\|I'f\|^2_{\mathcal{S}'_{\mathcal{N}}} = \mathbb{E}\big((I'f, W)^2\big) = \mathbb{E}\Big(\big(\sum_{i\geqslant 1} \sigma_i \xi_i c_i\big)^2\Big) = \sum_{i\geqslant 1} \sigma_i^2 c_i^2$$

Moreover, since $\|f\|_{\mathcal{S}} = \|c\|_{l_2}$, the continuity of I' is insured by the assumption that $\sum_i \sigma_i^2 < \infty$. Now one can easily check that the kernel space of the law of W is given by

$$\mathcal{K} = \Big\{ f_c = \sum_{i\geqslant 1} c_i e_i \ ; \ \sum_{i\geqslant 1} \big(\frac{c_i}{\sigma_i}\big)^2 < \infty \Big\},$$

endowed with inner product $(f_c, f_d)_{\mathcal{K}} = \sum_{i\geqslant 1} \frac{c_i d_i}{\sigma_i^2}$.

Reproducing Kernel Hilbert Spaces (RKHS): Note that if we now introduce the functions $\{\varphi_i\}_{i\geqslant 1}$ defined by $\varphi_i \overset{\text{def}}{=} \sigma_i e_i \in \mathcal{H}$, then we get

$$\mathcal{K} = \Big\{ f_\alpha = \sum_{i\geqslant 1} \alpha_i \varphi_i \ ; \ \|\alpha\|_{l_2} < \infty \Big\},$$

endowed with inner product $(f_\alpha, f_\beta)_{\mathcal{K}} = \langle \alpha, \beta \rangle_{l_2}$. If we consider for instance that $\mathcal{H} \subset L_{2,\mu}(\mathcal{X}; \mathbb{R})$ for some reference measure μ, and that $\{e_i\}_{i\geqslant 1}$ are orthonormal w.r.t. $L_{2,\mu}(\mathcal{X}; \mathbb{R})$, then $\mathcal{K}$ appears to be a RKHS that can be made fully explicit; its kernel is defined by $k(x,y) = \sum_{i=1}^{\infty} \sigma_i^2 e_i(x) e_i(y)$, and $(\sigma_i)_i$ and $(e_i)_i$ are trivially the eigenvalues and eigenfunctions of the integral operator $T_k : L_{2,\mu}(\mathcal{X}) \to L_{2,\mu}(\mathcal{X})$ defined by $(T_k(f))(x) = \int_{\mathcal{X}} k(x,y) f(y) d\mu(y)$.

Wavelet basis and Besov spaces: In this paragraph, we now apply this contruction to the case when the $\{e_i\}_{i\geqslant 1}$ are chosen to be a wavelet basis of functions defined on $\mathcal{X} = [0,1]$ with reference measure μ being the Lebesgue measure. Let e be the mother wavelet function, and let us write $e_{j,l}$ the ith element of the basis, where $j \in \mathbb{N}$ is a scale index and $l \in \{0, \dots, 2^j - 1\}$ is a position index, where we re-index all families indexed by i with the indices j, l. Let us define the coeficients $\{\sigma_i\}_{i\geqslant 1}$ to be exponentially decreasing with the scale index:

$$\sigma_{j,l} \stackrel{\text{def}}{=} 2^{-js} \text{ for all } j \geqslant 0 \text{ and } l \in \{0, \dots, 2^j - 1\},$$

where we introduced some positive real number s.

Now assume that for some $q \in \mathbb{N} \setminus \{0\}$ such that $q > s$, the mother wavelet function e belongs to $\mathcal{C}^q(\mathcal{X})$, the set of q-times continuously differentiable functions on $\mathcal{X}$, and admits q vanishing moments. In this case, the (homogeneous) Besov space $\mathcal{B}_{s,2,2}([0,1]^d)$ admits the following characterization (independent of the choice of the wavelets (Frazier and Jawerth, 1985, Bourdaud, 1995)):

$$\mathcal{B}_{s,2,2}(\mathcal{X};\mu) = \left\{ f \in L_{2,\mu}(\mathcal{X}) ; \ \|f\|^2_{s,2,2} \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \left[2^{2js} \sum_{l=0}^{2^j-1} |\langle f, e_{j,l} \rangle|^2 \right] < \infty \right\}$$

On the other hand, with the notations above, where in particular $\varphi_{j,l} = \sigma_{j,l}\varepsilon_{j,l}$, we deduce that the kernel space of the Gaussian object $W = \sum_{j,l} \xi_{j,l}\varphi_{j,l}$ (that we call a Scrambled wavelet), is the space

$$\mathcal{K} = \left\{ f_\alpha = \sum_{j,l} \alpha_{j,l}\varphi_{j,l} \ ; \ \sum_{j,l} \alpha^2_{j,l} < \infty \right\},$$

and a straightforward computation shows that $\|\alpha\|^2_{l_2} = \|f_\alpha\|^2_{s,2,2}$, and thus $\mathcal{K} = \mathcal{B}_{s,2,2}(\mathcal{X};\mu)$. Moreover, assuming that the mother wavelet is bounded by λ and has compact support $[0,1]$, then we have the property that

$$\sup_{x\in\mathcal{X}} \|\varphi(x)\|^2 \leqslant \frac{\lambda^2}{1 - 2^{-2s+1}}.$$

Note that a similar construction applies to the case when the orthonormal basis $\{e_i\}_{i\geqslant 1}$ is chosen to be a Fourier basis of functions, and the coeficients $\{\sigma_i\}_{i\geqslant 1}$ are chosen to be of the form $\sigma_i = i^{-s}$.

3.2.2 Gaussian objects defined by a Carleman expansion

We now no longer assume that the supporting space $\mathcal{S}$ is a Hilbert space. In this case, it is still possible to generate a Gaussian object with kernel space being a RKHS by resorting to Carleman operators.

A Carleman operator is a linear injective mapping $J : \mathcal{H} \mapsto \mathcal{S}$ (where $\mathcal{H}$ is a Hilbert space) such that $J(h)(t) = \int \Gamma_t(s)h(s)ds$ where $(\Gamma_t)_t$ is a collection of functions of $\mathcal{H}$. There

is a bijection between Carleman operators and the set of RKHSs (Canu et al., 2009, Saitoh, 1988). In particular, $J(\mathcal{H})$ is a RKHS.

A Gaussian object admitting $J(\mathcal{H})$ as a kernel space can be built as follows. By application of Lemma 6.2, we deduce that $\mathcal{K} = J(\mathcal{H})$ endowed with the inner product $\langle Jh_1, Jh_2\rangle_{\mathcal{K}} \stackrel{\text{def}}{=} \langle h_1, h_2\rangle_{\mathcal{H}}$ is the kernel space of $\mathcal{N}(0, JJ')$. Moreover, if we consider an orthonormal basis $\{e_i\}_{i\geqslant 1}$ of $\mathcal{H}$, an application of Lemma 6.2 shows that the functions $\{\varphi_i\}_{i\geqslant 1}$ defined by $\varphi_i = J(e_i)$ form an orthonormal basis of $J(\mathcal{H})$ and are such that the object $W = \sum_{i\geqslant 1} \xi_i \varphi$ is first a well-defined Gaussian object and then an expansion for the law $\mathcal{N}(0, JJ')$. We call this expansion a Carleman expansion. Note that this expansion is bottom-up whereas the Mercer expansion of a kernel via the spectral Theorem is top-down.

Cameron-Martin space We apply as an example this construction to the case of the Brownian motion and the Cameron-Martin space.

Let $\mathcal{S} = \mathcal{C}([0,1])$ be the space of continuous real-valued functions of the unit interval. Then $\mathcal{S}'$ is the set of signed measures and we can define the dual product by $(\nu, f) = \int_{[0,1]} f d\nu$. It is straightforward to check that the Brownian motion indexed by $[0,1]$ is a Gaussian object $W \in \mathcal{S}$, with $a \equiv 0$ and K defined by $(K\nu)(t) = \int_{[0,1]} \min(s,t)\nu(ds)$.

Kernel space. We consider the Hilbert space $\mathcal{H} = L_2([0,1])$ and define the mapping $J : \mathcal{H} \mapsto \mathcal{S}$ by

$$(Jh)(t) = \int_{[0,t]} h(s)ds\,;$$

simple computations show that $(J'\nu)(t) = \nu([t,1])$, $K = JJ'$ and that J is a Carleman operator. Therefore, the kernel space $\mathcal{K}$ is equal to $J(L_2([0,1]))$, or more explicitly

$$\mathcal{K} = \left\{k \in H^1([0,1]); k(0) = 0\right\},$$

where $H^1([0,1])$ is the Sobolev space of order 1.

Expansion of the Brownian motion. We build a Carleman expansion for the Brownian motion thanks to the Haar basis of $L^2([0,1])$, whose image by J defines an orthonormal basis of $\mathcal{K}$; The Haar basis is defined in a wavelet-way via the mother function $e(x) = \mathbb{I}_{[0,1/2[} - \mathbb{I}_{[1/2,1[}$ and the father function $e_0(x) = \mathbb{I}_{[0,1]}(x)$ as the functions $(e_{j,l})_{j,l\in\mathbb{N}}$ for any scale $j \geqslant 1$ and translation index $0 \leqslant l \leqslant 2^j - 1$ together with h_0, where

$$e_{j,l}(x) \stackrel{\text{def}}{=} 2^{j/2} e(2^j x - l)\,,$$

An orthonormal basis of the kernel space of the Brownian motion W and an expansion of W is thus obtained by

$$W = \sum_{j,l\geqslant 1} \xi_{j,l}\varphi_{j,l} + \xi_0\varphi_0,$$

$$\text{with } \varphi_{j,l}(x) = Je_{j,l}(x) = 2^{-j/2}\Lambda(2^j x - l) \text{ and } \varphi_0(x) = Je_0(x) = x\,,$$

where $\Lambda(x) = x\mathbb{I}_{[0,1/2[} + (1-x)\mathbb{I}_{[1/2,1[}$ is the mother hat function.

Bounded energy. Note that the rescaling factor inside $\varphi_{j,l}$ naturally appears as $2^{-j/2}$, and not as $2^{j/2}$ as usually defined in wavelet-like transformations. Note also that since the support of the mother function Λ is $[0,1]$, and also $\|\Lambda\|_\infty \leqslant 1/2$, then for any $x \in [0,1]^d$, for all j there exists at most one $l = l(x)$ such that $\varphi_{j,l}(x) \neq 0$, and we have the property that

$$\|\varphi(x)\|^2 = \sum_{j \geqslant 1} \varphi_{j,l(x)}(x)^2 \leqslant \sum_{j \geqslant 1} (2^{-j/2}\|\Lambda\|_\infty)^2 \leqslant \frac{1}{2}.$$

Remark 3 *This construction can be extended to the dimension $d > 1$ in at least two ways. Consider the space $\mathcal{S} = \mathcal{C}([0,1]^d)$, and the Hilbert space $\mathcal{H} = L_2([0,1]^d)$. Then if we define J to be the volume integral $(Jh)(t) = \int_{[0,t]} h(s)ds$ where $[0,t] \subset [0,1]^d$, this corresponds to the covariance operator defined by $(K\nu)(t) = \int_{[0,1]^d} \Pi_{i=1}^d \min(s_i, t_i)\nu(ds)$, i.e. to the Brownian sheet defined by tensorization of the Brownian motion. The corresponding kernel space in this case is thus $\mathcal{K} = J(L^2([0,1]^d))$, endowed with the norm $\|f\|_\mathcal{K} = \|\frac{\partial^d f}{\partial x_1 \dots \partial x_d}\|_{L^2([0,1]^d)}$. It corresponds to the* Cameron-Martin space *(Janson, 1997) of functions having a d-th order crossed (weak) derivative $\frac{\partial^d f}{\partial x_1 \dots \partial x_d}$ that belongs to $L^2([0,1]^d)$, vanishing on the "left" boundary (edges containing 0) of the unit d-dimensional cube. A second possible extension is to consider the isotropic Brownian sheet.*

3.3 A Johnson-Lindenstrauss lemma for Gaussian objects

In this section, we derive a version of the Johnson-Lindenstrauss' lemma that applies to the case of Gaussian objects.

The original Johnson-Lindenstrauss' lemma can be stated as follows ; its proof directly uses concentration inequalities (Cramer's large deviation Theorem from 1938) and may be found e.g. in Achlioptas (2003).

Lemma 6.3 *Let A be a $P \times F$ matrix of i.i.d. Gaussian $\mathcal{N}(0, 1/P)$ entries. Then for any vector α in $\mathbb{R}^F$, the random (with respect to the choice of the matrix A) variable $\|A\alpha\|^2$ concentrates around its expectation $\|\alpha\|^2$ when P is large: for $\varepsilon \in (0,1)$, we have*

$$\begin{aligned} \mathbb{P}\Big(\|A\alpha\|^2 \geqslant (1+\varepsilon)\|\alpha\|^2\Big) &\leqslant e^{-P(\varepsilon^2/4 - \varepsilon^3/6)} \\ \mathbb{P}\Big(\|A\alpha\|^2 \leqslant (1-\varepsilon)\|\alpha\|^2\Big) &\leqslant e^{-P(\varepsilon^2/4 - \varepsilon^3/6)} \end{aligned}$$

Remark 4 *Note that the Gaussianity is not mandatory here, and this is also true for other distributions, such as:*

- *Rademacher distributions, i.e. which takes values $\pm 1/\sqrt{P}$ with equal probability $1/2$,*
- *Distribution taking values $\pm\sqrt{3/P}$ with probability $1/6$ and 0 with probability $2/3$.*

This Lemma together with the measurability properties of Gaussian objects enable us to derive the following statement.

Lemma 6.4 *Let $(x_n)_{n\leqslant N}$ be N (deterministic) points of $\mathcal{X}$. Let $A : l_2(\mathbb{R}) \mapsto \mathbb{R}^P$ be the operator defined with i.i.d. Gaussian $\mathcal{N}(0, 1/P)$ variables $(A_{i,p})_{i\geqslant 1, p\leqslant P}$, such that for all $\alpha \in l_2(\mathbb{R})$, then*

$$(A\alpha)_p = \sum_{i\geqslant 1} \alpha_i A_{i,p} \, .$$

Let also define $\psi_p = \sum_{i\geqslant 1} A_{i,p}\varphi_i$, $f_\alpha = \sum_{i\geqslant 1} \alpha_i \varphi_i$ and $g_\beta = \sum_{p=1}^{P} \beta_p \psi_p$.

Then, A is well-defined and for all $P \geqslant 1$, for all $\varepsilon \in (0, 1)$, with probability larger than $1 - 4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ w.r.t. the Gaussian random variables,

$$\|f_\alpha - g_{A\alpha}\|_N^2 \leqslant \varepsilon^2 \|\alpha\|^2 \frac{1}{N} \sum_{n=1}^{N} \|\varphi(x_n)\|^2 \, ,$$

where we recall that by assumption $\varphi(x) = (\varphi_i(x))_{i\geqslant 1} \in l_2$ for all x.

This result is natural in view of concentration inequalities, since we have that for all $x \in \mathcal{X}$, the expectation $\mathbb{E}_{\mathcal{P}_\mathcal{G}}(g_{A\alpha}(x)) = f_\alpha(x)$ and the variance $\mathbb{V}_{\mathcal{P}_\mathcal{G}}(g_{A\alpha}(x)) = \frac{1}{P}(f_\alpha^2(x) + \|\alpha\|^2\|\varphi(x)\|^2)$. See the Appendix for the full proof.

Note also that a natural idea in order to derive generalization bounds would be to derive a similar result uniformly over $\mathcal{X}$ instead of a union bound over the samples. However, while such extension would be possible for finite dimensional spaces $\mathcal{F}$ (by resorting to covers) these kind of results are not possible in the general case, since $\mathcal{F}$ is typically big.

4 Regression with random subspaces

In this section, we describe the construction of the random subspace $\mathcal{G}_P \subset \mathcal{F}$ defined as the span of the random features $(\psi_p)_{p\leqslant P}$ generated from the initial features $(\varphi_i)_{i\geqslant 1}$. This method was originally described in Maillard and Munos (2009) for the case when $\mathcal{F}$ is of finite dimension, and we extend it here to the non-obvious case of infinite dimensional spaces $\mathcal{F}$, which relies on the fact that the randomly generated features $(\psi_p)_{p\leqslant P}$ are well-defined Gaussian objects.

The next subsection is devoted to the analysis of the approximation power of the random features space. We first give a survey of existing results on regression together with the standard hypothesis under which they hold in section 4.2, then we describe in section 4.4 an algorithm that builds the proposed regression function and provide excess risk bounds for this algorithm.

4.1 Construction of random subspaces

Assumption on initial features. In this chapter we assume that the set of features $(\varphi_i)_{i\geqslant 1}$ are continuous and satisfy the assumption that,

$$\sup_{x\in\mathcal{X}} \|\varphi(x)\|^2 < \infty, \text{ where } \|\varphi(x)\|^2 \stackrel{\text{def}}{=} \sum_{i\geqslant 1} \varphi_i(x)^2. \tag{6.3}$$

Note that all examples in Section 3 satisfy this condition.

Random features. The random subspace $\mathcal{G}_P$ is generated by building a set of P *random features* $(\psi_p)_{1\leqslant p\leqslant P}$ defined as linear combinations of the initial features $\{\varphi_i\}_{1\geqslant 1}$ weighted by random coefficients:

$$\psi_p(x) \stackrel{\text{def}}{=} \sum_{i\geqslant 1} A_{p,i}\varphi_i(x), \text{ for } 1 \leqslant p \leqslant P$$

where the (infinitely many) coefficients $A_{p,i}$ are drawn i.i.d. from a centered distribution with variance $1/P$. Here we explicitly choose a Gaussian distribution $\mathcal{N}(0, 1/P)$. Such a definition of the features ψ_p as an infinite sum of random variable is not obvious (this is an expansion of a Gaussian object) and we refer to the Section 3 for elements of theory about Gaussian objects and Lemma 6.2 for the expansion of a Gaussian object. It is shown that under Assumption (6.3), the random features are well defined. Actually, they are random samples of a centered Gaussian process indexed by the space $\mathcal{X}$ with covariance structure given by $\frac{1}{P}\langle\varphi(x), \varphi(x')\rangle$, where we used the notation $\langle u, v\rangle = \sum_i u_i v_i$ for two square-summable sequences u and v. Indeed, $\mathbb{E}_{A_p}[\psi_p(x)] = 0$, and

$$\mathrm{Cov}_{A_p}(\psi_p(x), \psi_p(x')) = \mathbb{E}_{A_p}[\psi_p(x)\psi_p(x')] = \frac{1}{P}\sum_{i\geqslant 1} \varphi_i(x)\varphi_i(x') = \frac{1}{P}\langle\varphi(x), \varphi(x')\rangle$$

The continuity of the initial features (φ_i) guarantees that there exists a continuous version of the process ψ_p which is thus a Gaussian process.

Random subspace. We finally define $\mathcal{G}_P \subset \mathcal{F}$ to be the (random) vector space spanned by those features, i.e.

$$\mathcal{G}_P \stackrel{\text{def}}{=} \{g_\beta(x) \stackrel{\text{def}}{=} \sum_{p=1}^{P} \beta_p\psi_p(x), \beta \in \mathbb{R}^P\}.$$

We now want to compute a high probability bound on the excess risk of an estimator built using the random space $\mathcal{G}_P$. To this aim, we first quickly review known results in regression and see what kind of estimator can be considered and what results can be applied. Then we compute a high probability bound on the approximation error of the considered random space w.r.t. to initial space $\mathcal{F}$. Finally, we combine both bounds in order to derive a bound on the excess risk of the proposed estimate.

4.2 Reminder of results on regression

Review of some results for regression For the sake of completeness, we now review other existing results in regression that may or may not apply in our setting. Indeed it seems natural to apply existing results for regression to the space $\mathcal{G}_P$. For that purpose, we focus on the randomness coming from the data points only, and not from the Gaussian entries. We will thus consider in this subsection only a space $\mathcal{G}$ that is the span over a *deterministic* set of P functions $\{\psi_p\}_{p\leqslant P}$, and for a convex subset $\Theta \subset \mathbb{R}^P$, we write

$$\mathcal{G}_\Theta = \{g_\theta \in \mathcal{G}; \theta \in \Theta\}\,.$$

Similarly, we write $g^\star = \arg\min_{g\in\mathcal{G}} R(g)$ and $g^\star_\Theta = \arg\min_{g\in\mathcal{G}_\Theta} R(g)$. Examples of well studied estimates are:

- $\widehat{g}^{ols} = \arg\min_{g\in\mathcal{G}} R_N(g)$, the ordinary least-squares (ols) estimate
- $\widehat{g}^{erm} = \arg\min_{g\in\mathcal{G}_\Theta} R_N(g)$ the empirical risk minimizer (erm), which coincides with the ols when $\Theta = \mathbb{R}^P$.
- $\widehat{g}^{ridge} = \arg\min_{g\in\mathcal{G}} R_N(g) + \lambda\|\theta\|$, $\widehat{g}^{lasso} = \arg\min_{g\in\mathcal{G}} R_N(g) + \lambda\|\theta\|_1$.

We also introduce for convenience g_B, the truncation at level $\pm B$ of some $g \in \mathcal{G}$ to be defined by $g_B(x) \stackrel{\text{def}}{=} T_B[g(x)]$, where $T_B(u) \stackrel{\text{def}}{=} \begin{cases} u & \text{if } |u| \leqslant B, \\ B\ \text{sign}(u) & \text{otherwise.} \end{cases}$

There are at least 9 different theorems that one may want to apply in our setting. Since those theorems hold under some assumptions, we list them now. Unfortunately, as we will see, these assumptions are usually slightly too strong to apply in our setting, and thus we will need to build our own analysis instead.

Assumptions Let us list the following assumptions.

- Noise assumptions: (for some constants B, B_1, σ, ξ)
  (N_1) $|Y| \leqslant B_1$,
  (N_2) $\sup_{x\in\mathcal{X}} \mathbb{E}(Y|X=x) \leqslant B$,
  (N_3) $\sup_{x\in\mathcal{X}} \mathbb{V}(Y|X=x) \leqslant \sigma^2$,
  (N_4) $\forall k \geqslant 3$ $\sup_{x\in\mathcal{X}} \mathbb{E}(|Y|^k|X=x) \leqslant \sigma^2 k!\xi^{k-2}$.
- Moment assumptions: (for some constants σ, a, M)
  (M_1) $\sup_{x\in\mathcal{X}} \mathbb{E}([Y - g^\star_\Theta(X)]^2|X=x) \leqslant \sigma^2$,
  (M_2) $\sup_{x\in\mathcal{X}} \mathbb{E}(\exp[a|Y - g^\star_\Theta(X)|]|X=x) \leqslant M$,
  (M_3) $\exists g_0 \in \mathcal{G}_\Theta$ $\sup_{x\in\mathcal{X}} \mathbb{E}(\exp[a|Y - g_0(X)|]|X=x) \leqslant M$.
- Function space assumptions for $\mathcal{G}$: (for some constant D)
  (G_1) $\sup_{g_1,g_2\in\mathcal{G}_\Theta} \|g_1 - g_2\|_\infty \leqslant D$,
  (G_2) $\exists g_0 \in \mathcal{G}_\Theta$, known, such that $\|g_0 - g^\star_\Theta\|_\infty \leqslant D$.

- Dictionary assumptions:
  (D_1) $L = \max_{1 \leqslant p \leqslant P} \|\psi_p\|_\infty < \infty$,
  (D_2) $L = \sup_{x \in \mathcal{X}} \|\psi(x)\|_2 < \infty$,
  (D_3) $\operatorname{esssup} \|\psi(X)\|_2 \leqslant L$,
  (D_4) $L = \inf_{\{\psi'_p\}_{p \leqslant P}} \sup_{\theta \in \mathbb{R}^d - \{0\}} \frac{\| \sum_{p=1}^P \theta_p \psi'_p \|_\infty}{\|\theta\|_\infty} < \infty$ where the infimum is over all orthonormal basis of $\mathcal{G}$ w.r.t. to $L_{2,\mathcal{P}_\mathcal{X}}(\mathcal{X};\mathbb{R})$.

- Orthogonality assumptions:
  (O_1) $\{\psi_p\}_{p \leqslant P}$ is an orthonormal basis of $\mathcal{G}$ w.r.t. to $L_{2,\mathcal{P}_\mathcal{X}}(\mathcal{X};\mathbb{R})$,
  (O_2) $\det(\Psi) > 0$, where $\Psi = \mathbb{E}(\psi(X)\psi(X)^T)$ is the Gram matrix.

- Parameter space assumptions:
  (P_1) $\sup_{\theta \in \Theta} \|\theta\|_\infty < \infty$,
  (P_2) $\|\theta^\star\|_1 \leqslant S$ where $\theta^\star$ is such that $g_{\theta^\star} = g^\star_\Theta$ and S is known,
  (P_3) $\sup_{\theta \in \Theta} \|\theta\|_2 \leqslant 1$.

Theorem 6.2 (Györfi et al. (2002)) *Let $\Theta = \mathbb{R}^P$. Under assumption (N_2) and (N_3), the truncated estimator $\widehat{g}_L = T_L(\widehat{g}^{ols})$ satisfies*

$$\mathbb{E}R(\widehat{g}_L) - R(f^{(reg)}) \leqslant 8[R(g^*) - R(f^{(reg)})] + \kappa \frac{(\sigma^2 \vee B^2) P \log(N)}{N}$$

where κ is some numerical constant and $f^{(reg)}(x) \overset{\text{def}}{=} \mathbb{E}(Y|X=x)$.

Theorem 6.3 (Catoni (2004)) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (M_3), (G_1) and (O_2), there exists constants $C_1, C_2 > 0$ (depending only on a, M and D) such that with probability $1-\delta$, provided that*

$$\left\{ g \in \mathcal{G} \,;\, R_N(g) \leqslant R_N(\widehat{g}^{ols}) + C_1 \frac{P}{N} \right\} \subset \mathcal{G}_\Theta \,,$$

then the ordinary least squares estimate satisfies

$$R(\widehat{g}^{ols}) - R(g^\star_\Theta) \leqslant C_2 \frac{P + \log(\delta^{-1}) + \log(\frac{det\,\widehat{\Psi}}{det\,\Psi})}{N} \,.$$

where $\widehat{\Psi} = \frac{1}{N} \sum_{i=1}^N \psi(X_i)\psi(X_i)^T$ is the empirical Gram matrix.

Theorem 6.4 (Audibert and Catoni (2010a) from Alquier (2008)) *Let $\Theta = \mathbb{R}^P$. Under assumption (N_1) and (G_2), there exists a randomized estimate $\widehat{g}$ that only depends on g_0, L, C, such that for all $\delta > 0$, with probability larger than $1-\delta$ w.r.t. all sources of randomness,*

$$R(\widehat{g}) - R(g^\star) \leqslant \kappa(B_1^2 + D^2) \frac{P \log(3\nu_{\min}^{-1}) + \log(\log(N)\delta^{-1})}{N} \,.$$

where κ does not depend on P and N, and $\nu_{\min}$ is the smallest eigenvalue of Ψ.

Theorem 6.5 (Koltchinskii (2006)) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (N_1), (D_3) and (P_3), $\widehat{g}^{erm}$ satisfies, for any $\delta > 0$ with probability higher than $1-\delta$,*

$$R(\widehat{g}^{erm}) - R(g_\Theta^\star) \leqslant \kappa(B_1 + L)^2 \frac{rank(\Psi) + \log(\delta^{-1})}{N}\,.$$

where κ is some constant.

Theorem 6.6 (Birgé and Massart (1998)) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (M_3), (G_1) and (D_4), for all $\delta > 0$ with probability higher than $1-\delta$,*

$$R(\widehat{g}^{erm}) - R(g_\Theta^\star) \leqslant \kappa(a^{-2} + D^2) \frac{P \log(2 + (L^2/N) \wedge (N/P)) + \log(\delta^{-1})}{N}\,.$$

where κ is some constant depending only on M.

Theorem 6.7 (Tsybakov (2003)) *Let $\Theta = \mathbb{R}^P$. Under assumption (N_2), (N_3) and (O_1), the projection estimate $\widehat{g}^{proj}$ satisfies*

$$\mathbb{E}(R(\widehat{g}^{proj})) - R(g^\star) \leqslant \frac{(\sigma^2 + B^2)P}{N}$$

Theorem 6.8 (Caponnetto and De Vito (2007)) *Under assumption (M_2) and (D_2), for all $\delta > 0$ for $\lambda = PL^2 \log^2(\delta^{-1})/N \leqslant \nu_{\min}$, with probability higher than $1-\delta$,*

$$R(\widehat{g}^{ridge}) - R(g_\Theta^\star) \leqslant \kappa(a^{-2} + \frac{\lambda L^2 \|\theta^\star\|^2 \log^2(\delta^{-1})}{\nu_{\min}}) \frac{P \log^2(\delta^{-1})}{N}\,.$$

where κ is some constant depending only on M.

Theorem 6.9 (Alquier and Lounici (2010)) *Let $\Theta = \mathbb{R}^P$ and define for all $\alpha \in (0,1)$ the prior $\pi_\alpha(J) = \frac{\alpha^{|J|}}{\sum_{i=0}^N \alpha^i} \binom{P}{|J|}^{-1}$ for all $J \subset 2^P$. Under assumption $(N_2), (N_3)$, (N_4), (D_1) and (P_2), by setting $\lambda = \frac{N}{2C}$ where*

$$C \stackrel{\text{def}}{=} \max\{64\sigma^2 + (2B + L(2S + \frac{1}{N}))^2, 64[\xi + 2B + L(2S + \frac{1}{N})]L(2S + \frac{1}{N})\}\,,$$

the randomized aggregate estimator $\widehat{g}$ defined in Alquier and Lounici (2010) based on prior π_α satisfies, for any $\delta > 0$ with probability higher than $1-\delta$,

$$R(\widehat{g}) - R(g_\Theta^\star) \leqslant C \frac{S^\star \log(\frac{(S+c)eNP}{\alpha S^\star}) + \log(2\delta^{-1}/(1-\alpha))}{N} + \frac{3L^2}{N^2},$$

where $S^\star = \|\theta^\star\|_0$.

Theorem 6.10 (Audibert and Catoni (2010a)) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (M_1), (G_1) and (P_1) so that one can define the uniform probability distribution over Θ, there exists a random estimator $\widehat{g}$ (drawn according to a Gibbs distribution $\widehat{\pi}$) that satisfies, with probability higher than $1-\delta$ w.r.t. all source of randomness,*

$$R(\widehat{g}) - R(g_\Theta^\star) \leqslant (2\sigma + D)^2 \frac{16.6P + 12.5\log(2\delta^{-1})}{N}\,.$$

Note that Theorem 6.2 and Theorem 6.7 provide a result in expectation only, which is not enough for our purpose, since we need high probability bounds on the excess risk in order to be able to handle the randomness of the space $\mathcal{G}_P$.

Assumptions satisfied by the random space $\mathcal{G}_P$ We now discuss the assumptions that are satisfied in our setting where $\mathcal{G}$ is a random space $\mathcal{G}_P$ built from the random features $\{\psi_p\}_{p\leqslant P}$, in terms of assumptions on the underlying initial space $\mathcal{F}$.

- The noise assumptions (N) do not concern $\mathcal{G}$.
- The moment assumptions (M) are not restrictive. By combining similar assumptions on $\mathcal{F}$, the results on approximation error of Section 4.3 can be shown to hold (with different constants).
- Assumptions (P) are generally too strong. For (P_1), the reason is that there is no high probability link between $\|A\alpha\|_\infty$ and $\|\alpha\|$ for usual norms. Now even if $\alpha^\star$ is sparse or has low l_1-norm, this does not imply this is the case for $\beta^\star = \arg\min_{\beta\in\mathbb{R}^P} R(g_\beta)$ or $A\alpha^\star$ in general, thus (P_2) cannot be assumed either. Finally (P_3) may be assumed in some case. Let us assume that we know that $\|\alpha^\star\|_2 \leqslant 1$. Then $\|A\alpha^\star\|_2 \leqslant 1+\varepsilon$ with high probability, thus it is enough to consider the space $\mathcal{G}_P(\Theta)$ with parameter space $\Theta = \{\beta; \|\beta\|_2 \leqslant (1+\varepsilon)\}$, and thus $A\alpha^\star \in \Theta$ with high probability.
- Assumptions (G) are strong assumptions. The reason is that it is difficult to relate the vector coefficient $\beta^\star$ or even $A\alpha^\star$ to the vector coefficient $\alpha^\star$ of $f^\star = f_{\alpha^\star}$ in l_∞ norm. Thus even if we know some f_0 close to $f^\star$ in l_∞-norm, this does not imply that we can build a function g_0 close to $g^\star = g_{\beta^\star}$.
- Assumptions (D) will not be valid a.s. w.r.t. the law of the Gaussian variables. The assumptions (D_1) and (D_4) are difficult to satisfy since they concern $\|.\|_\infty$. For assumption (D_2) and (D_3), we have the property that for each x, $\|\psi(x)\|_2^2$ is close to $\|\varphi(x)\|_2^2$ with high probability. However, we need here a uniform result over $x \in \mathcal{X}$ which seems difficult to get since the space $\mathcal{F}$ is actually big (not of finite dimension).
- Assumptions (O), which are typically strong assumptions for specific features φ appear to be almost satisfied. The reason is due to the covariance structure of the random

features. Indeed whatever the distribution $\mathcal{P}_{\mathcal{X}}$ (independent of $\mathcal{P}_{\mathcal{G}}$), we have that $\langle \psi_p, \psi_q \rangle$ concentrates around

$$\mathbb{E}_{\mathcal{P}_{\mathcal{G}}} \langle \psi_p, \psi_q \rangle = \frac{1}{P} \| \sum_{i \geqslant 1} \varphi_i \|^2_{\mathcal{P}_{\mathcal{X}}} \delta_{p,q} \,,$$

where $\delta_{p,q}$ is the Kronecker symbol between p and q. Thus the orthogonality assumption is satisfied with high probability. Note that the knowledge of $\mathcal{P}_{\mathcal{X}}$ is still needed in order to rescale the features and obtain orthonormality. Similar argument shows that (O_2) is also valid.

As a consequence, only Theorems 6.2 and 6.7 would apply safely, but unfortunately these Theorems do not give results in high probability.

In the next two sections, we derive similar results but in high probability with assumptions that corresponds to our setting. We provide a hand-made Theorem that makes use of the technique introduced in Györfi et al. (2002) and that can be applied without too restrictive assumptions, although not being optimal in terms of constant and logarithmic factors.

4.3 Approximation power of random spaces

We assume that $f^\star = f_{\alpha^\star} \in \mathcal{F}$.

Theorem 6.11 (Approximation error with deterministic design) *For all $P \geqslant 1$, for all $\delta \in (0,1)$ there exists an event of $\mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta$ such that on this event,*

$$\inf_{g \in \mathcal{G}_P} \|f^\star - g\|_N^2 \leqslant 12 \frac{\log(4N/\delta)}{P} \|\alpha^\star\|^2 \frac{1}{N} \sum_{n=1}^{N} \|\varphi(x_n)\|^2 \,.$$

Theorem 6.12 (Approximation error with random design) *Under assumption (N_2) then for all $P \geqslant 1$, for all $\delta \in (0,1)$, the following bound holds with $\mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta$:*

$$\inf_{g \in \mathcal{G}_P} \|f^\star - T_B(g)\|^2_{\mathcal{P}_{\mathcal{X}}} \leqslant 25 \frac{\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}{P} \Big(1 + \frac{1}{2} \log \big(\frac{P \log(8P/\gamma^2\delta)}{18 \gamma^2 \delta} \big) \Big) \,,$$

where $\gamma \stackrel{\text{def}}{=} \frac{1}{B} \|\alpha^\star\| \sup_x \|\varphi(x)\|$ and T_B is the truncation operator at level B.

The result is not trivial because of the randomness of the space $\mathcal{G}_P$. Thus in order to keep the explanation simple, the proof (detailed in the Appendix) makes use of Hoeffding's Lemma only, which relies on the bounded assumption of the features (which can be seen either as a nice assumption, since it is simple and easy to check, or as a too strong assumption for some cases). Note that this result can be further refined by making use, for instance, of moment assumptions on the feature space instead.

4.4 Excess risk of random spaces

In this section, we analyze the excess risk of the random projection method. Thus for a proposed random estimate $\widehat{g}$, we are interested in bounding $R(\widehat{g}) - R(f^\star)$ in high probability with respect to any source of randomness.

4.4.1 Regression algorithm.

From now on we consider the estimate $\widehat{g}$ to be the least-squares estimate $g_{\widehat{\beta}} \in \mathcal{G}_P$ that is the function in $\mathcal{G}_P$ with minimal empirical error, i.e.

$$g_{\widehat{\beta}} = \arg\min_{g_\beta \in \mathcal{G}_P} R_N(g_\beta), \tag{6.4}$$

and is the solution of a least-squares regression problem, i.e. $\widehat{\beta} = \Psi^\dagger Y \in \mathbb{R}^P$ with matrix-wise notations, where $Y \in \mathbb{R}^N$ is here the vector of observations (not to be confused with the random variable Y that shares the same notation), Ψ is the $N \times P$-matrix composed of the elements: $\Psi_{n,p} \stackrel{\text{def}}{=} \psi_p(x_n)$, and $\Psi^\dagger$ is the Moore-Penrose pseudo-inverse[2] of Ψ. The final prediction function $\widehat{g}(x)$ is the truncation (at level $\pm B$) of $g_{\widehat{\beta}}$, i.e. $\widehat{g}(x) \stackrel{\text{def}}{=} T_B[g_{\widehat{\beta}}(x)]$.

In the next subsection, we provide excess risk bounds w.r.t. $f^\star$ in $\mathcal{G}_P$.

4.4.2 Regression with deterministic design

Theorem 6.13 *Under assumption (N_1), then for all $P \geqslant 1$, for all $\delta \in (0,1)$ there exists an event of $\mathcal{P}_\mathcal{Y} \times \mathcal{P}_\mathcal{G}$-probability higher than $1-\delta$ such that on this event, the excess risk of the estimator $g_{\widehat{\beta}}$ is bounded as*

$$\|f^\star - g_{\widehat{\beta}}\|_N^2 \leqslant \frac{12\log(8N/\delta)}{P}\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2 + \kappa B_1^2 \frac{P + \log(2/\delta)}{N},$$

for some numerical constant $\kappa > 0$.

Note that from this theorem, we deduce (without further assumptions on the features $\{\varphi_i\}_{i\geqslant 1}$) that for instance for the choice $P = \frac{\sqrt{N}}{\log(N/\delta)}$ then

$$\|f^\star - g_{\widehat{\beta}}\|_N^2 \leqslant \kappa' \Big[\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2 \sqrt{\frac{\log(N/\delta)}{N}} + \frac{\log(1/\delta)}{N} \Big]$$

for some positive constant κ'. Note also that whenever an upper-bound on the square terms $\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2$ is known, this can be used in the definition of P in order to improve this bound.

[2] In the full rank case when $N \geqslant P$, $\Psi^\dagger = (\Psi^T\Psi)^{-1}\Psi^T$

4.4.3 Regression with random design

In the regression problem with random design, the analysis of the excess risk of a given method is not straightforward, since the assumptions to apply standard techniques may not be satisfied without further knowledge on the structure of the features. In a general case, we can use the techniques introduced in Györfi et al. (2002), which yields to the following (not optimal) result:

Theorem 6.14 *Under assumption* (N_1) *and* (N_2)*, provided that* $N\log(N) \geqslant \frac{4}{P}$ *(thus whenever* $\min(N,P) \geqslant 2$*), then with* $\mathcal{P}_{\mathcal{G}} \times \mathcal{P}$*-probability at least* $1-\delta$*,*

$$R(T_B(g_{\widehat{\beta}})) - R(f^\star) \leqslant \kappa\Big[\frac{\log(12N/\delta)}{P}\|\alpha^\star\|^2 \sup_{x\in\mathcal{X}}\|\varphi(x)\|^2 + \max\{B_1^2, B^2\}\frac{P + P\log(N) + \log(3/\delta)}{N}\Big]\,.$$

for some positive constant κ*.*

Let us now provide some intuition about the proof of this result. We first start by explaining what does not work. A natural idea in order to derive this result would be to consider the following decomposition:

$$R(T_B(g_{\widehat{\beta}})) - R(f^\star) \leqslant [R(T_B(g_B^\star)) - R(f^\star)] + [R(T_B(g_{\widehat{\beta}})) - R(T_B(g_B^\star))]\,.$$

where $g_B^\star \in \arg\min_{g\in\mathcal{G}} R(T_B(g)) - R(f^\star)\,.$

Indeed the first term is controlled with high $\mathcal{P}_{\mathcal{G}}$-probability by Theorem 6.12, and since $R(g_{\widehat{\beta}}) - R(g_B^\star) \leqslant R(g_{\widehat{\beta}}) - R(g^\star)$, the second term is controlled for each fixed $\omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}$ with high $\mathcal{P}$-probability by standard Theorems for regression, provided that we can relate $R(T_B(g_{\widehat{\beta}})) - R(T_B(g_B^\star))$ to $R(g_{\widehat{\beta}}) - R(g_B^\star)$. Thus by doing the same careful analysis of the events involved, this should lead to the desired result.

However, the difficulty lies first in ensuring that the conditions of application of standard Theorems are satisfied with high $\mathcal{P}_{\mathcal{G}}$-probability and then in relating the excess risk of the truncated function to that of the non-truncated ones, since it is not true in general that $R(T_B(g_{\widehat{\beta}})) - R(T_B(g_B^\star)) \leqslant R(g_{\widehat{\beta}}) - R(g_B^\star)$. Thus we resort to a different decomposition in order to derive our results. The sketch of proof of Theorem 6.14 actually consists in applying the following lemma.

Lemma 6.5 *The following decomposition holds for all* $C > 0$

$$\|T_B(g_{\widehat{\beta}}) - f^\star\|_{\mathcal{P}_{\mathcal{X}}}^2 \;\leqslant\; C\|f^\star - g_{\tilde{\beta}}\|_N^2 + C\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 + \sup_{g\in\mathcal{G}}\left(\|f^\star - T_B(g)\|_{\mathcal{P}_{\mathcal{X}}}^2 - C\|f^\star - T_B(g)\|_N^2\right),$$

where $g_{\tilde{\beta}} = \Pi_{\|.\|_N}(f^\star, \mathcal{G})$ *and* $g_{\widehat{\beta}} = \Pi_{\|.\|_N}(Y, \mathcal{G})$ *are the projections of the target function* $f^\star$ *and observation* Y *onto the random linear space* $\mathcal{G}$ *with respect to the empirical norm* $\|.\|_N$*.*

We then call the first term $\|f^\star - g_{\tilde{\beta}}\|_N^2$ an approximation error term, the second $\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2$ a noise error term and the third one $\sup_{g\in\mathcal{G}}\left(\|f^\star - T_B(g)\|_{\mathcal{P}_{\mathcal{X}}}^2 - C\|f^\star - T_B(g)\|_N^2\right)$ an estimation error term.

In order to prove Theorem 6.14, we then control each of these terms: We apply Lemma 6.11 to the first term, Lemma 6.6 below to the second term and finally Theorem 11.2 of Györfi et al. (2002) to the last term with $C = 8$, and the result follows by gathering all the bounds.

Let us now explain the contribution to each of the three terms in details.

Approximation error term The first term, $\|f^\star - g_{\tilde{\beta}}\|_N^2$, is an approximation error term in empirical norm, it contains the number of projections as well as the norm of the target function. This term plays the role of the approximation term that exists for regression with penalization by a factor $\lambda\|f\|^2$. This term is controlled by application of Theorem 6.11 conditionally on the random samples, and then w.r.t. all source of randomness by independence of the Gaussian random variables with the random samples.

Noise error term The second term, $\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2$, is an error term due to the observation noise η. This term classically decreases at speed $\frac{D\sigma^2}{N}$ where σ^2 is the variance of the noise and D is related to the log entropy of the space of function $\mathcal{G}$ considered. Without any more assumption, we only know that this is a linear space of dimension P, so this term finally behaves like $\frac{P\sigma^2}{N}$, but note that this dependency with P may be improved depending on the knowledge about the functions ψ (for instance, if $\mathcal{G}$ is included in a Sobolev space of order s, we would have $P^{1/2s}$ instead).

Lemma 6.6 *Under assumption (N_1), then for each realization of the Gaussian variables, with $\mathcal{P}$-probability higher than $1-\delta$, the following holds true:*

$$\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 \leqslant 6B_1^2 \frac{1616P + 200\log(6/\delta) + \log(3/\delta)}{N}.$$

Note that we may consider different assumptions on the noise term. Here we considered only that the noise is upper-bounded as $\|\eta\|_\infty \leqslant B_1$, but another possible assumption is that the noise has finite variance σ^2 or that the tail of the distribution of the noise behaves nicely, e.g., that $\|\eta\|_{\psi_\alpha} \leqslant B$, where ψ_α is the Orlicz norm or order α, with $\alpha = 1$ or 2.

Estimation error term The third term, $\sup_{g\in\mathcal{G}_P}(\|f^\star - T_B(g)\|_{\mathcal{P}_\mathcal{X}}^2 - \|f^\star - T_B(g)\|_N^2)$, is an estimation error term due to finiteness of the data. This term also depends on the log entropy of the space of functions, thus the same remark applies to the dependency with P as for the noise error term. We bound the third term by applying Theorem 11.2 of Györfi et al. (2002) to the class of functions $\mathcal{G}^0 = \{f^\star - T_B(g), g \in \mathcal{G}_P\}$, for fixed random Gaussian variables. Note that for all $f \in \mathcal{G}^0$, $\|f\|_\infty \leqslant 2B$. The precise result of Györfi et al. (2002) is the following :

Theorem 6.15 *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ bounded in absolute value by B. Let $\varepsilon > 0$. Then*

$$\mathbb{P}(\sup_{f\in\mathcal{F}} \|f\|_{\mathcal{P}_\mathcal{X}} - 2\|f\|_N > \varepsilon) \leqslant 3\mathbb{E}(\mathcal{N}(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{F}, \|.\|_{2N}))\exp(-\frac{N\varepsilon^2}{288B^2}).$$

We now have the following lemma whose proof is given in the Appendix:

Lemma 6.7 *Assuming that $N\log(N) \geqslant \frac{4}{P}$, then for each realization of the Gaussian variables, with $\mathcal{P}$-probability higher than $1-\delta$, the following holds true:*

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|^2_{\mathcal{P}_\mathcal{X}} - 8\|f^\star - T_B(g)\|^2_N \leqslant (24B)^2 \frac{4\log(3/\delta) + 2P\log(N)}{N}\,.$$

5 Discussion

5.1 Non-linear approximation

In the work (Barron et al., 2008), the authors provide excess risk bounds for greedy algorithms (i.e. in a non-linear approximation setting). The precise result they derive in their Theorem 3.1 is reported now, using the notations of section 4.2:

Theorem 6.16 (Barron et al. (2008)) *Consider spaces $\{\mathcal{G}_P\}_{P\geqslant 1}$ generated respectively by the span of features $\{e_p\}_{p\leqslant P}$ with increasing dimension P (thus $\Theta = \mathbb{R}^P$ for each P). For each $\mathcal{G}_P$ we compute a corresponding greedy empirical estimate $\widehat{g}_P \in \mathcal{G}_P$ provided by some algorithm (see Barron et al. (2008)), then we define $\widehat{P} = \arg\min \|y - T_{B_1}\widehat{f}_P\|^2_N + \kappa\frac{P\log(N)}{N}$ for some constant κ, and finally define $\widehat{g} = T_{B_1}(\widehat{g}_{\widehat{P}})$, and fix some P_0.*

Under assumption (N_1), there exists κ_0 depending only on B_1 and a where $P_0 = \llcorner N^a \lrcorner$ such that if $\kappa \geqslant \kappa_0$, then for all $P > 0$ and for all functions g_θ in $\mathcal{G}_{P_0}$, the estimator $\widehat{g}$ satisfies

$$\mathbb{E}R(\widehat{g}) - R(f^{(reg)}) \leqslant 2[R(g_\theta) - R(f^{(reg)})] + 8\frac{\|\theta\|_1^2}{P} + C\frac{P\log N}{N}\,,$$

where the constant C only depends on κ, B_1 and a.

The bound is thus similar to that of Theorem 6.14 in Section 4.4. One difference is that this bound contains the l_1 norm of the coefficients θ^* while the l_2 norm of the coefficients $\alpha^\star$ appears in our setting. We leave as an open question to understand whether this difference is a consequence of the non-linear aspect of their approximation or if it results from the different assumptions made about the approximation spaces, in terms of rate of decrease of the coefficients.

The main difference is actually about the tractability of the proposed estimator, since it relies on greedy estimation that is computationally heavy while on the other hand, random projection is cheap (see Subsection 5.4).

5.2 Adaptivity

Randomization enables to define approximation spaces such that the approximation error (either in expectation or in high probability on the choice of the random space) is controlled, whatever the measure $\mathcal{P}$ used to assess the performance is (which is specially interesting in the

regression setting where $\mathcal{P}$ is unknown). As mentioned in the introduction, because the choice of the subspace $\mathcal{G}_P$ within which we perform the least-squares estimate is random, we avoid (with high probability) degenerated situations where the target function $f^\star$ cannot be well approximated with functions in $\mathcal{G}_P$. Indeed, in methods that consider a given (deterministic) finite-dimensional space $\mathcal{G}$ of the big space $\mathcal{F}$ (like linear approximation using a predefined set of wavelets), it is often possible to find a target function $f^\star$ such that $\inf_{g\in\mathcal{G}_P} \|f^\star - g\|_N$ is large, whereas using this method, the random choice of $\mathcal{G}_P$ implies that for any $f^\star \in \mathcal{F}$, the approximation error $\inf_{g\in\mathcal{G}_P} \|f^\star - g\|_N$ can be controlled (by the first term of the bound (6.2)) in high probability. We now illustrate this property on a simple example.

Example Let us consider a very peaky (a spot) distribution $\mathcal{P}$. Regular linear approximation, say with wavelets (see e.g. DeVore (1997)), will most probably miss the specific characteristics of $f^\star$ at the spot, since the first wavelets have large support. On the contrary, the random features $\{\psi_p\}_{p\leqslant P}$ that are functions that contain (random combinations of) all wavelets, will be able to detect correlations between the data and some high frequency wavelets, and thus discover relevant features of $f^\star$ at the spot. This is illustrated in the numerical experiment below.

Here $\mathcal{P}$ is a very peaky Gaussian distribution and $f^\star$ is a 1-dimensional periodic function. We consider as initial features $(\varphi_i)_{i\geqslant 1}$ the set of hat functions defined in Section 3.2.2. Figure 5.2 shows the target function $f^\star$, the distribution $\mathcal{P}$, and the data $(x_n, y_n)_{1\leqslant n\leqslant 100}$ (left plots). The middle plots represents the least-squares estimate $\widehat{g}$ using $P = 40$ scrambled objects $(\psi_p)_{1\leqslant p\leqslant 40}$ (here Brownian motions). The right plots shows the least-squares estimate using the initial features $(\varphi_i)_{1\leqslant i\leqslant 40}$. The top figures represent a high level view of the whole domain $[0, 1]$. No method is able to learn $f^\star$ on the whole space (this is normal since the available data are only generated from a peaky distribution). The bottom figures shows a zoom $[0.45, 0.51]$ around the data. Least-squares regression using scrambled objects is able to learn the structure of $f^\star$ in terms of the measure $\mathcal{P}$.

5.3 Other related work

In Rahimi and Recht (2008, 2007), the authors consider, for a given parameterized function $\Phi : \mathcal{X} \times \Theta \to \mathbb{R}$ bounded by 1, and a probability measure μ over Θ, the space $\mathcal{F}$ of functions $f(x) = \int_\Theta \alpha(\theta)\Phi(x, \theta)d\theta$ such that $\|f\|_\mu = \sup_\theta |\frac{\alpha(\theta)}{\mu(\theta)}| < \infty$. They show that this is a dense subset of the RKHS with kernel $k(x, y) = \int_\Theta \mu(\theta)\Phi(x, \theta)\Phi(y, \theta)d\theta$, and that if $f \in \mathcal{F}$, then with high probability over $(\theta_p)_{p\leqslant P} \overset{i.i.d}{\sim} \mu$, there exist coefficients $(c_p)_{p\leqslant P}$ such that $\widehat{f}(x) = \sum_{p=1}^P c_p\Phi(x, \theta_p)$ satisfies $\|\widehat{f} - f\|_2^2 \leqslant O(\frac{\|f\|_\mu}{\sqrt{P}})$. The method is analogous to the construction of the empirical estimates $g_{A\alpha} \in \mathcal{G}_P$ of function $f_\alpha \in \mathcal{K}$ in our setting. Indeed we may formally identify $\Phi(x, \theta_p)$ with $\psi_p(x) = \sum_i A_{p,i}\varphi_i(x)$, θ_p with the sequence $(A_{p,i})_i$, and the distribution μ with the distribution of this infinite sequence. However, in our setting we do not require the condition $\sup_{x,\theta} \Phi(x, \theta) \leqslant 1$ to hold and the fact that Θ is a set of infinite sequences makes the identification tedious without the Gaussian random functions

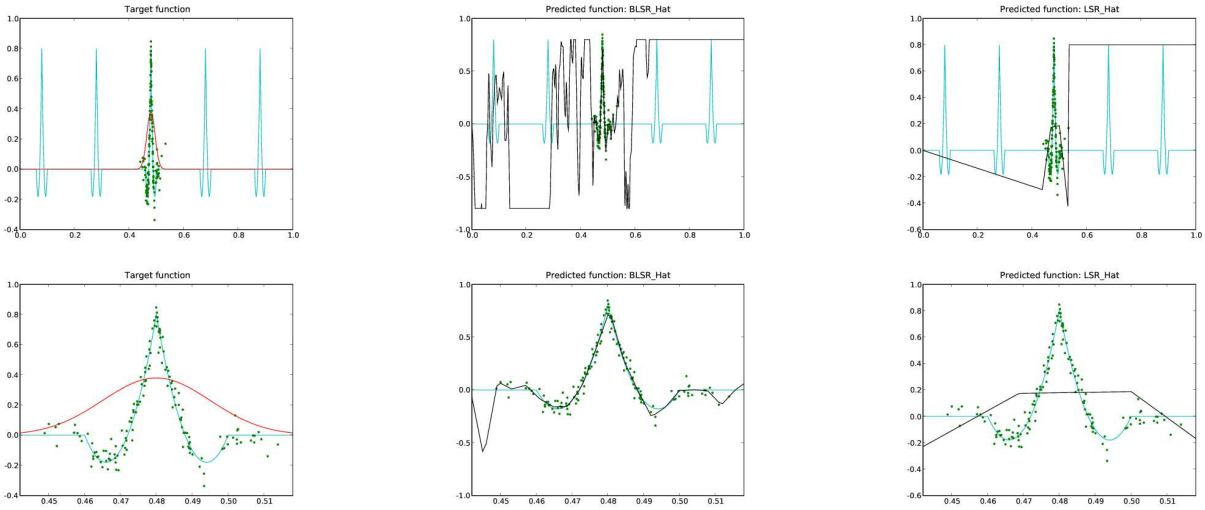


Figure 6.3: LS estimate of $f^\star$ using $N = 100$ data generated from a peaky distribution $\mathcal{P}$ (left plots), using 40 Brownian motions (ψ_p) (middle plots) and 40 hat functions (φ_i) (right plots). The bottom row shows a zoom around the data.

theory used here. Anyway, we believe that this link provides a better mutual understanding of both approaches (i.e. Rahimi and Recht (2008) and this work).

5.4 Tractability

In practice, in order to build the least-squares estimate, one needs to compute the values of the random features $(\psi_p)_{1\leqslant p\leqslant P}$ at the data points $(x_n)_{1\leqslant n\leqslant N}$, i.e. the matrix $\Psi = (\psi_p(x_n))_{p\leqslant P, n\leqslant N}$. Moreover, due to finite memory and precision of computers, numerical implementations can only handle a finite number F of initial features $(\varphi_i)_{1\leqslant i\leqslant F}$.

Approximation error Using a finite F introduces an additional approximation (squared) error term in the final excess risk bounds. This additional error (due to the numerical approximation) is of order $O(F^{-\frac{2s}{d}})$ for a wavelet basis adapted to $H^s([0,1]^d)$ and can be made arbitrarily small, e.g. $o(N^{-1/2})$, whenever the depth of the wavelet dyadic-tree is bigger than $\frac{\log N}{d}$. Our main concern is thus about efficient computation.

Numerical complexity In Maillard and Munos (2009) it was mentioned that the computation of Ψ, which makes use of the random matrix $A = (A_{p,i})_{p\leqslant P, i\leqslant F}$, has a complexity $O(FPN)$.

In the multi-resolution schemes described now, provided that the mother function has compact support (such as the hat functions), we can significantly speed up the computation of the matrix Ψ by using a *tree-based lazy expansion*, i.e. where the expansion of the random

features $(\psi_p)_{p\leqslant P}$ is built only when needed for the evaluation at the points $(x_n)_n$. Note that in the specific case of wavelets, we can even think to combine random projection with tools like fast wavelet transform which would be even faster (which we do not do here for simplicity).

Example: Consider the example of the scrambled wavelets. In dimension 1, using a wavelet dyadic-tree of depth H (i.e. $F = 2^{H+1}$), the numerical cost for computing Ψ is $O(HPN)$ (using one tree per random feature). Now, in dimension d the classical extension of one-dimensional wavelets uses a family of 2^d-1 wavelets, thus requires 2^d-1 trees each one having 2^{dH} nodes. While the resulting number of initial features F is of order $2^{d(H+1)}$, thanks to the lazy evaluation (notice that one never computes all the initial features), one needs to expand at most one path of length H per training point, and the resulting complexity to compute Ψ is $O(2^dHPN)$. Thus the method is linear with N and reduces the amount of computation by an exponential factor (from 2^{dH} to 2^dH).

Note that one may alternatively use the so-called sparse-grids instead of wavelet trees, which have been introduced by Griebel and Zenger (see Zenger (1990), Bungartz and Griebel (2004)). The main result is that one can reduce significantly the total number of features to $F = O(2^H H^d)$ (while preserving a good approximation for sufficiently smooth functions). Similar lazy evaluation techniques can be applied to sparse-grids.

Thus, using $P = O(\sqrt{N})$ random features, we deduce that the complexity of building the matrix Ψ is at most $O(2^d N^{3/2} \log N)$. Then in order to solve the least squares system, one has to compute $\Psi^T\Psi$, that has cost at most $O(P^2N)$, and then solve the system by inversion, which has numerical cost $O(P^{2.376})$ by Coppersmith and Winograd (1987). Thus, with $P = O(\sqrt{N})$, the overall cost of the algorithm is $O(2^d N^{3/2} \log N + N^2)$ without fancy computations designed for random matrices, and the numerical complexity to make a new prediction is $O(2^d N^{1/2} \log(N))$.

Acknowledgements

The authors want to thank *Pierre Chainais* and *Olivier Degris* for interesting pointers to the literature in image processing and applied reinforcement learning.

6 Technical details

6.1 Proof of Lemma 6.4

Proof: **Step 1.** First, we derive a result similar to Lemma 6.3 that holds for dot products, by polarisation of the Euclidean norm. The precise statement for our purpose is the following one.

Lemma 6.8 *Let A be a $P \times F$ matrix of i.i.d. elements drawn from one of the previously defined distributions. Let $(u_n)_{1\leqslant n\leqslant N}$ and v be $N+1$ vectors of $\mathbb{R}^F$.*

Then for any $\varepsilon \in (0,1)$, with probability at least $1 - 4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$, simultaneoulsy for all $n \leqslant N$,

$$|Au_n \cdot Av - u_n \cdot v| \leqslant \varepsilon \|u_n\| \, \|v\| \, .$$

We apply Lemma 6.3 to any couple of vectors $u+w$ and $u-w$, where u and w are vectors of norm 1. By polarisation, we have that

$$\begin{aligned} 4Au \cdot Aw &= \|Au + Aw\|^2 - \|Au - Aw\|^2 \\ &\leqslant (1+\varepsilon)\|u+w\|^2 - (1-\varepsilon)\|u-w\|^2 \\ &= 4u \cdot w + \varepsilon(\|u+w\|^2 + \|u-w\|^2) \\ &= 4u \cdot w + 2\varepsilon(\|u\|^2 + \|w\|^2) = 4u \cdot w + 4\varepsilon \end{aligned}$$

fails with probability $2e^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ (we applied the previous lemma twice at line 2).

Thus for each $n \leqslant N$, we have with same probability:

$$Au_n \cdot Av \leqslant u_n \cdot v + \varepsilon \|u_n\| \, \|v\|.$$

Now the symmetric inequality holds with the same probability, and using a union bound for considering all $(u_n)_{n \leqslant N}$, we have that

$$|Au_n \cdot Av - u_n \cdot v| \leqslant \varepsilon \|u_n\| \, \|v\|,$$

holds for all $n \leqslant N$, with probability $1 - 4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$.

Step 2. We now extend this Lemma to the case of infinite sequences. This is made possible thanks to the measurability properties of Gaussian Objects. Indeed, for any given F, Lemma 6.8 applies to the two truncated sequences $\overline{\alpha}_F = (\alpha_1, \ldots, \alpha_F)$ and $\overline{\varphi}_F(x_n) = (\varphi_1(x_n), \ldots, \varphi_F(x_n))$; this gives that for all n simultaneoulsy,

$$|\sum_{i=1}^{F} \alpha_i \varphi_i(x_n) - \frac{1}{P} \sum_{p=1}^{P} \big(\sum_{i=1}^{F} \xi_{i,p} \alpha_i \big) \big(\sum_{i=1}^{F} \xi_{i,p} \varphi_i(x_n) \big)| \leqslant \varepsilon \, \|\overline{\alpha}_F\| \, \|\overline{\varphi}_F(x_n)\|$$

happens with probability higher than $1 - 4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$, where we introduced $\xi_{i,p} \overset{\text{def}}{=} \sqrt{P} A_{i,p} \sim \mathcal{N}(0,1)$ in order to avoid confusion with the section on Gaussian objects. Now by the assumption that $\alpha \in l_2(\mathbb{R})$ and $\varphi(x) \in l_2(\mathbb{R})$ for all x, then the Gaussian objects $\sum_{i=1}^{\infty} \xi_{i,p} \alpha_i$ and $\sum_{i=1}^{\infty} \xi_{i,p} \varphi_i(x_n)$ are well-defined square integrable random variables. Thus, taking the limit of the above inequality when F tends to ∞ yields that with same probability, for all $n \leqslant N$

$$|f_\alpha(x_n) - g_{A\alpha}(x_n)| \leqslant \varepsilon \, \|\alpha\| \, \|\varphi(x_n)\| \, .$$

□

6.2 Proof of Lemma 6.6

Proof: We can bound the noise term $\|g_{\tilde\beta} - g_{\widehat\beta}\|_N^2$ using a simple Chernoff bound together with a chaining argument. Indeed, by definition of $g_{\tilde\beta}$ and $g_{\widehat\beta}$, if we introduce the noise vector η defined by $\eta = Y - f$, we have

$$\begin{aligned}
\|g_{\tilde\beta} - g_{\widehat\beta}\|_N^2 &= \langle g_{\tilde\beta} - g_{\widehat\beta}, \eta\rangle_N \\
&= \frac{1}{N}\sum_{i=1}^N \eta_i (g_{\tilde\beta} - g_{\widehat\beta})(X_i) \\
&\leqslant \Big(\sup_{g\in\mathcal{G}} \frac{\frac{1}{N}\sum_{i=1}^N \eta_i g(X_i)}{\|g\|_N}\Big)\|g_{\tilde\beta} - g_{\widehat\beta}\|_N \\
&\leqslant \Big(\sup_{g\in\mathcal{G}} \frac{\frac{1}{N}\sum_{i=1}^N \eta_i g(X_i)}{\|g\|_N}\Big)^2 .
\end{aligned}$$

Thus, we focus on the set $\mathcal{G}^1 = \{g \in \mathcal{G}; \|g\|_N = 1\}$. Note that since $\mathcal{G}^1$ is a sphere in a space of dimension P, its ε-packing number in empirical norm is bounded above by $\mathcal{M}(\varepsilon, \mathcal{G}^1, \|.\|_N) \leqslant \mathcal{N}(\varepsilon/2, \mathcal{G}^1, \|.\|_N) \leqslant \mathcal{N}(\varepsilon/2, \{g \in \mathcal{G}; \|g\|_N \leqslant 1\}, \|.\|_N) \leqslant (\frac{4}{\varepsilon} + 1)^P \leqslant \max(\frac{5}{\varepsilon}, 5)^P$, where $\mathcal{N}$ refers here to the covering number.

We now introduce for convenience the following notation, for fixed Gaussian random variables and data points $(X_i)_{i=1..n}$:

$$\rho(t) \stackrel{\text{def}}{=} \mathbb{P}_Y\big(\exists g \in \mathcal{G} \frac{\frac{1}{N}\sum_{i=1}^N \eta_i g(X_i)}{\|g\|_N} > t\big) = \mathbb{P}_Y\big(\exists g \in \mathcal{G}^1 \frac{1}{N}\sum_{i=1}^N \eta_i g(X_i) > t\big).$$

For $j = 0...\infty$, let us consider ε_j-packings C_j of $\mathcal{G}^1$ for the empirical norm $\|.\|_N$, with $C_0 = g_0$, such that C_{j+1} is a refinement of C_j and $\varepsilon_j \leqslant \varepsilon_{j-1}$. Then for a given $g \in \mathcal{G}^1$, we define $g_j = \Pi(g, C_j)$ the projection of g into C_j, for the norm $\|g\|_N$. Thus, $g - g_0 = (g - g_J) + \sum_{j=1}^J (g_j - g_{j-1})$. Note that since by definition of $\mathcal{G}^1$ we have $\|g - g_0\|_N \leqslant 2$, we need to consider $\varepsilon_0 \geqslant 2$.

Thus if we now introduce real numbers γ and $(\gamma_j)_{j\geqslant 1}$ such that $\sum_{j=1}^J \gamma_j \leqslant \gamma$, then we have

$$\begin{aligned}
\rho(\gamma t_1 + t_2 + t_3) &\leqslant \mathbb{P}\Big(\exists g \in \mathcal{G}^1 \frac{1}{N}\sum_{i=1}^N \eta_i (g - g_0)(X_i) > \gamma t_1 + t_2\Big) + \exp(-\frac{t_3^2 N}{2B_1^2}) \\
&\leqslant \mathbb{P}\Big(g \in \exists \mathcal{G}^1 \frac{1}{N}\sum_{i=1}^N \eta_i (g - g_J)(X_i) + \\
&\qquad \sum_{j=1}^J \frac{1}{N}\sum_{i=1}^N \eta_i (g_j - g_{j-1})(X_i) \geqslant \sum_{j=1}^J \gamma_j t_1 + t_2\Big) + \exp(-\frac{t_3^2 N}{2B_1^2}),
\end{aligned}$$

where we applied Hoeffding's inequality in the first line. We further have:

$$\begin{aligned}\rho(\gamma t_1 + t_2 + t_3) &\leqslant \sum_{j=1}^{J} \mathbb{P}\Big(\exists g \in \mathcal{G}^1 \frac{1}{N}\sum_{i=1}^{N} \eta_i (g_j - g_{j-1})(X_i) > t_1\gamma_j\Big) \\ &\quad + \exp(-\frac{t_2^2 N}{2B_1^2\varepsilon_J^2}) + \exp(-t_3^2 N 2B_1^2) \\ &\leqslant \mathbb{E}\sum_{j=1}^{J} \mathcal{M}(\varepsilon_j, \mathcal{G}^1, \|.\|_N)\mathcal{M}(\varepsilon_{j-1}, \mathcal{G}^1, \|.\|_N)\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} \eta_i (g_j - g_{j-1})(X_i) > t_1\gamma_j\Big) \\ &\quad + \exp(-\frac{t_2^2 N}{2B_1^2\varepsilon_J^2}) + \exp(-\frac{t_3^2 N}{2B_1^2})\,.\end{aligned}$$

Now, note that since $\varepsilon_j \leqslant \varepsilon_{j-1}$, then $\mathcal{M}(\varepsilon_{j-1}, \mathcal{G}^1, \|.\|_N) \leqslant \mathcal{M}(\varepsilon_j, \mathcal{G}^1, \|.\|_N)$. Note also that $\|g_j - g_{j-1}\|_N \leqslant \eta_j$ since C_j is a refinement of C_{j-1}. Finally, we can bound the packing number by $\mathcal{M}(\varepsilon_j, \mathcal{G}^1, \|.\|_N) \leqslant N_j = \max(\frac{5}{\varepsilon_j}, 5)^P$ where P is the dimension of $\mathcal{G}$. Thus we deduce that:

$$\rho(\gamma t_1 + t_2 + t_3) \leqslant \sum_{j=1}^{J} N_j^2 \exp(-\frac{t_1^2 N\gamma_j^2}{2B_1^2\varepsilon_j^2}) + \exp(-\frac{t^2 N}{2B_1^2\varepsilon_J^2}) + \exp(-\frac{t_3^2 N}{2B_1^2}).$$

Now, we define $\gamma_j = \frac{2\varepsilon_j B_1}{t_1}\sqrt{\frac{2\log(N_j)}{N}}$, $t_2 = B_1\varepsilon_J\sqrt{\frac{2\log(1/\delta_2)}{N}}$ and $t_3 = B_1\sqrt{\frac{2\log(1/\delta_3)}{N}}$, for some $\delta_2, \delta_3 \in (0, 1]$. Thus, we get:

$$\rho(\eta t_1 + t_2 + t_3) \leqslant \sum_{j=1}^{J} \frac{1}{N_j^2} + \delta_2 + \delta_3.$$

Thus, it remains to define ε_j. Since $N_j = \max(\frac{5}{\varepsilon_j}, 5)^P$, we define the covering radius ε_j to be $\varepsilon_j = 2^{-j}5\delta_1^{1/2P}(2^{2P} - 1)^{1/2P}$ for some $\delta_1 \in (0, 1]$, which entails that $\sum_{j=1}^{J} \frac{1}{N_j^2} \leqslant \delta_1$. Now since $\varepsilon_j \to 0$ when $j \to \infty$, we can make the sum goes to infinity. We deduce that:

$$\rho(\gamma t_1 + B_1\sqrt{\frac{2\log(1/\delta_3)}{N}}) \leqslant \delta_1 + \delta_2 + \delta_3.$$

Now, in order to bound the term $\gamma t_1 + t_2 + t_3$,we look at the following term:

$$\begin{aligned}
\gamma t_1 &= 2\sum_{j=1}^{\infty} \varepsilon_j B_1 \sqrt{\frac{2\log(N_j)}{N}} \\
&\leqslant \frac{20B_1}{\sqrt{N}} \sum_{j=1}^{\infty} 2^{-j} \sqrt{2jP\log(2) + \log(1/\delta_1) - \log(2^{2P}-1)} \\
&\leqslant \frac{20B_1}{\sqrt{N}} \sum_{j=1}^{\infty} 2^{-j} \sqrt{2(j-1)P\log(2) + \log(2/\delta_1)} \\
&\leqslant \frac{20B_1}{\sqrt{N}} \Big(\sum_{j=1}^{\infty} 2^{-j} \sqrt{2(j-1)P\log(2)} + \sqrt{\log(2/\delta_1)} \Big) \\
&\leqslant \frac{20B_1}{\sqrt{N}} \Big((1+\sqrt{2})\sqrt{2P\log(2)} + \sqrt{\log(2/\delta_1)} \Big).
\end{aligned}$$

where we use the fact that $\sum_{j=1}^{\infty} 2^{-j} \leqslant 1$, and that $\sum_{j=1}^{\infty} 2^{-j}\sqrt{(j-1)} \leqslant 1+\sqrt{2}$.

Using the inequalities $\sqrt{a}+\sqrt{b}+\sqrt{c} \leqslant \sqrt{3(a+b+c)}$, we thus deduce the following bound:

$$\begin{aligned}
\gamma t_1 + t_2 + t_3 &\leqslant \frac{B_1}{\sqrt{N}} \Big(20(1+\sqrt{2})\sqrt{2P\log(2)} + 20\sqrt{\log(2/\delta_1)} + \sqrt{2\log(1/\delta_3)} \Big) \\
&\leqslant \frac{\sqrt{6}B_1}{\sqrt{N}} \sqrt{400\log(2)(1+\sqrt{2})^2 P + 200\log(2/\delta_1) + \log(1/\delta_3)}.
\end{aligned}$$

Thus, by setting $\delta_1 = \delta_2 = \delta_3 = \delta/3$, we deduce that with $\mathcal{P}$-probability higher than $1-\delta$,

$$\sup_{g\in\mathcal{G}_P} \frac{\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i g(X_i)}{\|g\|_N} \leqslant \frac{B_1\sqrt{6}}{\sqrt{N}} \sqrt{400\log(2)(1+\sqrt{2})^2 P + 200\log(6/\delta) + \log(3/\delta)}\,.$$

□

6.3 Proof of Lemma 6.7

Proof: Indeed, let us introduce the space of functions $\mathcal{G}^0 = \{f^\star - T_B(g), g \in \mathcal{G}_P\}$. Then we have for $g \in \mathcal{G}^0$, $\|g\|_N \leqslant \|g\|_\infty \leqslant 2B$. Thus Theorem 11.2 of Györfi et al. (2002) gives the following bound:

$$\mathbb{P}\Big(\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_\mathcal{X}} - 2\|f^\star - T_B(g)\|_N > \varepsilon\Big) \leqslant 3\mathbb{E}\big(\mathcal{N}(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{G}^0, \|.\|_{2N})\big) \exp\Big(-\frac{N\varepsilon^2}{288(2B)^2}\Big).$$

Then, since $\mathcal{G}^0 = f^\star + T_B(\mathcal{G}_P)$, we bound the entropy number by:

$$\mathcal{N}(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{G}^0, \|.\|_{2N}) \leqslant \mathcal{N}(\frac{\sqrt{2}}{24}\varepsilon, T_B(\mathcal{G}_P), \|.\|_{2N}) \leqslant (\frac{2(2B).24}{\sqrt{2}\varepsilon} + 1)^P.$$

Thus we deduce that if $\varepsilon \geqslant \frac{24.4B}{\sqrt{2}}u$, then with probability higher than $1-\delta$ w.r.t $\mathbb{P}$, for fixed random Gaussian variables,

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_\mathcal{X}} - 2\|f^\star - T_B(g)\|_N \leqslant \varepsilon = 24B\sqrt{\log(3/\delta) + P\log(\frac{1}{u}+1)}\sqrt{\frac{2}{N}}.$$

Thus,we consider $u = \frac{1}{N-1}$, and deduce that, provided that $N\log(N) \geqslant \frac{4}{P}$, then with probability higher than $1-\delta$ w.r.t $\mathbb{P}$, for fixed random Gaussian variables (i.e. conditionally on them),

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_\mathcal{X}} - 2\|f^\star - T_B(g)\|_N \leqslant 24B\sqrt{\frac{2\log(3/\delta) + P\log(N)}{N}}.$$

Thus, we deduce that on this event, for all $g \in \mathcal{G}_P$

$$\begin{aligned}\|f^\star - T_B(g)\|^2_{\mathcal{P}_\mathcal{X}} &\leqslant (2\|f^\star - T_B(g)\|_N + 24B\sqrt{\frac{2\log(3/\delta) + P\log(N)}{N}})^2 \\ &\leqslant 8\|f^\star - T_B(g)\|^2_N + (24B)^2\frac{4\log(3/\delta) + 2P\log(N)}{N}.\end{aligned}$$

This gives the following upper bound, that holds with probability higher than $1-\delta$:

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|^2_{\mathcal{P}_\mathcal{X}} - 8\|f^\star - T_B(g)\|^2_N \leqslant (24B)^2\frac{4\log(3/\delta) + 2P\log(N)}{N}$$

□

6.4 Proof of Theorem 6.11

Proof: Since by assumption $f^\star = f_{\alpha^\star}$ for some $\alpha^\star$, we have by direct application of Lemma 6.4

$$\inf_{g\in\mathcal{G}} \|f^\star - g\|^2_N \leqslant \|f_{\alpha^\star} - g_{A\alpha^\star}\|^2_N .$$

Now let us define for some $N \geqslant 1$ the quantity $\varepsilon = \varepsilon_N(\delta)$ that appears in Lemma 6.4, such that

$$\frac{\log(4N/\delta)}{P} = \frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}.$$

Thus, since $\varepsilon \in (0,1)$, this means in particular that we have

$$\frac{\varepsilon^2}{3} \leqslant 4\frac{\log(4N/\delta)}{P} \leqslant \varepsilon^2 .$$

□

6.5 Proof of Theorem 6.12

Proof: By assumption, we consider that $f^\star \in \mathcal{F}$. Thus there exists a sequence $\alpha^\star \in \mathbb{R}^{\mathbb{N}}$ such that one can write:

$$f^\star \;=\; f_{\alpha^\star} \;=\; \sum_{i\geqslant 1} \alpha_i^\star \varphi_i \,,$$

Thus we consider in the sequel one such $\alpha^\star$. This enables to derive the following upper bound:

$$\inf_{g\in\mathcal{G}} \|f^\star - T_L(g)\|^2_{\mathcal{P}_\mathcal{X}} \leqslant \|f_{\alpha^*} - T_L(g_{A\alpha^*})\|^2_{\mathcal{P}_\mathcal{X}}.$$

where we applied the gaussian operator A to the sequence $\alpha^\star$.

Step 1. Applying Johnson-Lindenstrauss' Lemma. Let us introduce m ghost samples $(X'_j)_{j\leqslant m}$ i.i.d. according to $\mathcal{P}_\mathcal{X}$, and thus consider the following associated norm

$$\|f_{\alpha^*} - T_L(g_{A\alpha^*})\|^2_m = \frac{1}{m}\sum_{j=1}^m (f_{\alpha^*} - T_L(g_{A\alpha^*}))^2(X'_j)\,.$$

We now make explicit the probability spaces corresponding to the different sources of randomness. Consider the probability space defined over the product sample space $\Omega_\mathcal{X}\times\Omega_\mathcal{G}$, where $\Omega_\mathcal{X}$ consists of all the possible realizations of J states $X'_1,\ldots,X'_m$ drawn i.i.d. from $\mathcal{P}_\mathcal{X}$, and $\Omega_\mathcal{G}$ is the set of all possible realizations of the random elements $(A_{p,i})_{1\leqslant p\leqslant P, i\geqslant 1}$ (which define the random feature space $\mathcal{G}_P$).

Let us fix some $\omega_\mathcal{G}\in\Omega_\mathcal{G}$ (which defines the random subspace $\mathcal{G}_P(\omega_\mathcal{G})$). Since for all j, we have that $(f_{\alpha^*} - T_L(g_{A\alpha^*}))^2(X'_j) \in [0, 4L^2]$ $\mathcal{P}_\mathcal{X}$-a.s., then Hoeffding's inequality applies; we deduce that there exists an event $\Omega_\mathcal{X}(\omega_\mathcal{G})$ of $\mathcal{P}_\mathcal{X}$-probability higher than $1-\delta_\mathcal{X}$ such that on this event

$$\|f_{\alpha^*} - T_L(g_{A\alpha^*})\|^2_{\mathcal{P}_\mathcal{X}} \leqslant \|f_{\alpha^*} - T_L(g_{A\alpha^*})\|^2_m + (2L)^2\sqrt{\frac{\log(1/\delta)}{2m}}.$$

Now by independence between the Gaussian random variables and the sample, the same inequality is valid on the event

$$\Omega_1 = \{\omega_\mathcal{X}\times\omega_\mathcal{G}; \omega_\mathcal{G}\in\Omega_\mathcal{G}, \omega_\mathcal{X}\in\Omega_\mathcal{X}(\omega_\mathcal{G})\}$$

and this event has $\mathcal{P}_\mathcal{X}\times\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta_\mathcal{X}$.

In order to bound the first term of the right hand side of this inequality, we first notice that since $\|f_{\alpha^*}\|_\infty \leqslant L$, then

$$\|f_{\alpha^*} - T_L(g_{A\alpha^*})\|^2_m \;\;\leqslant\;\; \|f_{\alpha^*} - g_{A\alpha^*}\|^2_m\,,$$

then for some fixed $\omega_\mathcal{X}\in\Omega_\mathcal{X}$, that last term is bounded by $\varepsilon^2\|\alpha^\star\|^2\sup_x\|\varphi(x)\|^2$ on an event $\Omega_\mathcal{G}(\omega_\mathcal{X})$ of $\mathcal{P}_\mathcal{G}$-probability higher than $1-4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ by application of Lemma 6.4.

Thus still by independence, the same inequality is valid on the event

$$\Omega_2 = \{(\omega_\mathcal{X},\omega_\mathcal{G}); \omega_\mathcal{X}\in\Omega_\mathcal{X}, \omega_\mathcal{G}\in\Omega_\mathcal{G}(\omega_\mathcal{X})\}$$

and this event has $\mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{G}}$-probability higher than $1 - 4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$.

Thus, we deduce, by a union bound that for all $\varepsilon \in (0,1)$ and $m \geqslant 1$ there exists an event $\Omega_1 \cap \Omega_2$ of $\mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{G}}$-probability higher than $1 - \delta_{\mathcal{X}} - 4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ such that on this event,

$$\inf_{g \in \mathcal{G}} \|f^\star - T_L(g)\|^2_{\mathcal{P}_{\mathcal{X}}} \leqslant \varepsilon^2 \|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2 + (2L)^2 \sqrt{\frac{\log(1/\delta)}{2m}} \,.$$

Finally in order to get a bound in high $\mathcal{P}_{\mathcal{G}}$-probability only, we introduce for any $\omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}$ the event $\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}}) \stackrel{\text{def}}{=} \{\omega_{\mathcal{X}} \in \Omega_{\mathcal{X}}; (\omega_{\mathcal{X}}, \omega_{\mathcal{G}}) \in \Omega_1 \times \Omega_2\}$ and then define for all $\lambda > 0$ the event

$$\Lambda \stackrel{\text{def}}{=} \{\omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}; \mathbb{P}_{\mathcal{X}}(\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}})) \geqslant 1 - \lambda\} \,.$$

Using this notation, we deduce that for all $\omega_{\mathcal{G}} \in \Lambda$, the following bound holds

$$\begin{aligned}
\inf_{g \in \mathcal{G}_P(\omega_{\mathcal{G}})} \|f^\star - T_L(g)\|^2_{\mathcal{P}_{\mathcal{X}}} &\leqslant \int_{\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}})} \inf_{g \in \mathcal{G}_P(\omega_{\mathcal{G}})} \|f^\star - T_L(g)\|^2_{\mathcal{P}_{\mathcal{X}}} d\omega_{\mathcal{X}} \\
&\quad + \int_{\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}})^c} \inf_{g \in \mathcal{G}_P(\omega_{\mathcal{G}})} \|f^\star - T_L(g)\|^2_{\mathcal{P}_{\mathcal{X}}} d\omega_{\mathcal{X}} \\
&\leqslant \varepsilon^2 \|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2 + (2L)^2 \sqrt{\frac{\log(1/\delta)}{2m}} + (2L)^2 \lambda \,.
\end{aligned}$$

Moreover, since $\mathbb{P}_{\mathcal{X}\times\mathcal{G}}(\Omega_1 \cap \Omega_2) \geqslant 1 - \delta_{\mathcal{X}} - 4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ and on the other side

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}\times\mathcal{G}}(\Omega_1 \cap \Omega_2) &= \int_{\Omega_{\mathcal{G}}} \mathbb{P}_{\mathcal{X}}(\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}})) d\omega_{\mathcal{G}} \\
&\leqslant \int_{\Omega_{\mathcal{G}}} \mathbb{I}_{\mathbb{P}_{\mathcal{X}}(\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}})) \geqslant 1-\lambda} d\omega_{\mathcal{G}} + (1-\lambda) \int_{\Omega_{\mathcal{G}}} \mathbb{I}_{\mathbb{P}_{\mathcal{X}}(\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}})) < 1-\lambda} d\omega_{\mathcal{G}} \\
&\leqslant \mathbb{P}_{\mathcal{G}}(\Lambda) + (1-\lambda)(1 - \mathbb{P}_{\mathcal{G}}(\Lambda)) \,,
\end{aligned}$$

then we deduce that $\mathbb{P}_{\mathcal{G}}(\Lambda)) \geqslant 1 - \frac{\delta_{\mathcal{X}} + 4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}}{\lambda}$.

Step 2. Tuning the parameters ε. Now let us introduce $\delta_{\mathcal{G}}$ and define for some $m \geqslant 1$ the quantity $\varepsilon = \varepsilon_m(\delta_{\mathcal{G}})$ such that

$$\frac{\log(4m/\delta\mathcal{G})}{P} = \frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6} \,.$$

Thus, since $\varepsilon \in (0,1)$, this means in particular that we have

$$\frac{\varepsilon^2}{3} \leqslant 4\frac{\log(4m/\delta_{\mathcal{G}})}{P} \leqslant \varepsilon^2 \,.$$

Now by rewriting the bound using $\delta = \frac{\delta_{\mathcal{X}} + \delta_{\mathcal{G}}}{\lambda}$, we deduce that for all δ, for all m and λ, there exists an event of $\mathcal{P}_{\mathcal{G}}$-probability higher than $1 - \delta$ such that

$$\inf_{g \in \mathcal{G}} \|f^\star - T_L(g)\|^2_{\mathcal{P}_{\mathcal{X}}} \leqslant 12\frac{\log(\frac{8m}{\lambda\delta})}{P} \|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2 + (2L)^2 \Big(\sqrt{\frac{\log(\frac{2}{\lambda\delta})}{2m}} + \lambda \Big) \,.$$

Step 3. Optimizing over λ and m. Now, it remains to optimize the free parameter m and λ in this last bound; the optimal value for m is given by

$$m_{opt} = \frac{P^2 L^4 \log(\frac{2}{\lambda\delta})}{72\|\alpha\|^4 \sup_x \|\varphi(x)\|^4},$$

and the corresponding bound is thus

$$\inf_{g\in\mathcal{G}} \|f^\star - T_L(g)\|^2_{\mathcal{P}_\mathcal{X}} \leqslant 24\frac{\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}{P}\Big(1 + \log\big(\frac{PL^2\sqrt{\log(2/\lambda\delta)/\lambda\delta}}{3\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}\big)\Big) + (2L)^2\lambda\,.$$

Now one can take $\lambda \stackrel{\text{def}}{=} \frac{\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}{(2L)^2 P}$ and deduce the final bound. □

6.6 Proof of Theorem 6.13

Proof: We make use of the following decomposition:

$$\|f^\star - g_{\widehat{\beta}}\|^2_N \leqslant \|f^\star - g_{\tilde{\beta}}\|^2_N + \|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|^2_N\,,$$

and introduce the sets $\Omega_\mathcal{G}$ that consists of all possible realizations of the random elements $(A_{p,i})_{1\leqslant p\leqslant P, i\geqslant 1}$, and $\Omega_\mathcal{Y}$ that corresponds to the observation variables Y.

High $\mathcal{P}_\mathcal{Y}\times\mathcal{P}_\mathcal{G}$-probability bound. We again make explicit the probability spaces. For the first term on right hand side, an application of Theorem 6.11 ensures that there exists an event $\Omega'_\mathcal{G} \subset \Omega_\mathcal{G}$ of $\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta$ such that for all $\omega_\mathcal{G} \in \Omega'_\mathcal{G}$,

$$\|f^\star - g_{\tilde{\beta}}\|^2_N \leqslant 12\frac{\log(4N/\delta)}{P}\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2\,.$$

Since no random variable Y appears in this term, this is also true on the event

$$\Omega_1 \stackrel{\text{def}}{=} \{(\omega_\mathcal{Y}, \omega_\mathcal{G}) \in \Omega_\mathcal{Y}\times\Omega_\mathcal{G}; \omega_\mathcal{G} \in \Omega'_\mathcal{G}\}\,,$$

and Ω_1 has $\mathcal{P}_\mathcal{Y}\times\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta$.

For the second term, let us fix some $\omega_\mathcal{G} \in \Omega_\mathcal{G}$. Then Lemma 6.6 below shows that there exists an event $\Omega_\mathcal{Y}(\omega_\mathcal{G}) \subset \Omega_\mathcal{G}$ of $\mathcal{P}_\mathcal{Y}$-probability higher than $1-\delta'$ such that for all $\omega_\mathcal{Y} \in \Omega_\mathcal{Y}(\omega_\mathcal{G})$,

$$\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|^2_N \leqslant \kappa B^2 \frac{P + \log(1/\delta')}{N}\,,$$

for some numerical constant $\kappa > 0$. Thus by independence of the noise term with the Gaussian variables, we deduce that a similar bound holds on the event

$$\Omega_2 \stackrel{\text{def}}{=} \{(\omega_\mathcal{Y}, \omega_\mathcal{G}) \in \Omega_\mathcal{Y}\times\Omega_\mathcal{G}; \omega_\mathcal{Y} \in \Omega_\mathcal{Y}(\omega_\mathcal{G})\}\,,$$

and that Ω_2 has $\mathcal{P}_\mathcal{Y}\times\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta'$. Thus, we conclude by a simple union bound in order te get a result in high $\mathcal{P}_\mathcal{Y}\times\mathcal{P}_\mathcal{G}$-probability. □

6.7 Proof of Theorem 6.14

Proof:

Similarly to the proof of Theorem 6.12, we introduce the sets $\Omega_{\mathcal{X}}, \Omega_\eta$ and $\Omega_{\mathcal{G}}$ that consist of all possible realizations of the input, noise and Gaussian random variables. We then define $\Omega = \Omega_{\mathcal{X}} \times \Omega_\eta \times \Omega_{\mathcal{G}}$.

Step 1. High $\mathcal{P} \times \mathcal{P}_{\mathcal{G}}$-probability bound. In order to get a high probability bound, we use the decomposition given by Lemma 6.5. Now let us consider some fixed $\omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}$. One can apply Lemma 6.6 and Lemma 6.7 below for the noise and estimation term.

Thus when $N \log(N) \geqslant \frac{4}{P}$, there exists an event $\Omega_1(\omega_{\mathcal{G}})$ of $\mathcal{P}$-probability higher than $1-\delta_1$ and an event $\Omega_2(\omega_{\mathcal{G}})$ of $\mathcal{P}$-probability higher than $1-\delta_2$ such that for all $(\omega_{\mathcal{X}}, \omega_\eta) \in \Omega_1(\omega_{\mathcal{G}})$ we have

$$\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 \leqslant 6B^2 \frac{(1616P + 200\log(6/\delta) + \log(3/\delta))}{N},$$

and for all $(\omega_{\mathcal{X}}, \omega_\eta) \in \Omega_2(\omega_{\mathcal{G}})$ we have

$$\sup_{g \in \mathcal{G}_P} \|f^\star - T_L(g)\|_{\mathcal{P}_{\mathcal{X}}}^2 - 8\|f^\star - T_L(g)\|_N^2 \leqslant (24L)^2 \frac{4\log(3/\delta) + 2P\log(N)}{N}.$$

On the other hand, by application of Theorem 6.11, for any given $(\omega_{\mathcal{X}}, \omega_\eta)$, there exists an event $\Omega_{\mathcal{G}}(\omega_{\mathcal{X}}, \omega_\eta) \subset \Omega_{\mathcal{G}}$ of $\mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta_3$ such that on this event

$$\|f^\star - g_{\tilde{\beta}}\|_N^2 \leqslant 12 \frac{\log(4N/\delta_3)}{P} \|\alpha^\star\|^2 \sup_{x \in \mathcal{X}} \|\varphi(x)\|^2.$$

Thus by independence of the noise, data points and Gaussian variables, the three previous inequalities are valid respectively on the events

$$\begin{aligned}
\Omega_1 &= \{(\omega_{\mathcal{X}}, \omega_\eta, \omega_{\mathcal{G}}) \in \Omega; (\omega_{\mathcal{X}}, \omega_\eta) \in \Omega_1(\omega_{\mathcal{G}})\} \\
\Omega_2 &= \{(\omega_{\mathcal{X}}, \omega_\eta, \omega_{\mathcal{G}}) \in \Omega; (\omega_{\mathcal{X}}, \omega_\eta) \in \Omega_2(\omega_{\mathcal{G}})\} \\
\Omega_3 &= \{(\omega_{\mathcal{X}}, \omega_\eta, \omega_{\mathcal{G}}) \in \Omega; \omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}(\omega_{\mathcal{X}})\}
\end{aligned}$$

Moreover Ω_1 has $\mathcal{P} \times \mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta_1$, Ω_2 has $\mathcal{P} \times \mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta_2$, and Ω_1 has $\mathcal{P} \times \mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta_3$. We thus conclude by a simple union bound, and then by some cosmetic simplifications introducing some constant κ. □

Chapter 7

Brownian Sensing for the Recovery of a Sparse Function.

The previous chapter showed the benefit of using random matrices in order to address the problem of predicting as well as a target unknown function. In this chapter, we now turn to a related problem called recovery, where the goal is to recover the parameter of decomposition of the unknown function. Note that this problem is generally harder since recovering such parameters implies a law prediction error as well.

We consider the problem of recovering the parameter of decomposition $\alpha \in \mathbb{R}^K$ of a sparse function f (i.e. the number of non-zero entries of α is small compared to the number K of features) on a family of functions $\{\varphi_k\}_{1 \leqslant k \leqslant K}$ given noisy evaluations of f at a set of well-chosen sampling points. We introduce an additional randomization process, called Brownian sensing, based on the computation of stochastic integrals, which produces a Gaussian sensing matrix, for which good recovery properties are proven, independently on the number of sampling points N, even when the features are arbitrarily non-orthogonal. Under the assumption that f is Hölder continuous with exponent at least $1/2$, we provide an estimate $\widehat{\alpha}$ of the parameter such that $\|\alpha - \widehat{\alpha}\|_2 = O(\|\eta\|_2/\sqrt{N})$, where η is the observation noise. The method uses a set of sampling points uniformly distributed along a one-dimensional curve selected according to the features. We report numerical experiments illustrating our method.

This is a joint work with *Alexandra Carpentier*, with whom it is pleasant to work. A paper corresponding to this chapter has been accepted for publication in the 25*th conference on advances in Neural Information Processing Systems (NIPS 2011).*

Contents

1 Introduction

We consider the problem of sensing an unknown function $f : \mathcal{X} \to \mathbb{R}$ (where $\mathcal{X} \subset \mathbb{R}^d$), where f belongs to span of a large set of (known) features $\{\varphi_k\}_{1 \leqslant k \leqslant K}$ of $L_2(\mathcal{X})$:

$$f(x) = \sum_{k=1}^{K} \alpha_k \varphi_k(x),$$

where $\alpha \in \mathbb{R}^K$ is the unknown parameter, and is assumed to be S-sparse, i.e. that we have $\|\alpha\|_0 \stackrel{\text{def}}{=} |\{i : \alpha_k \neq 0\}| \leqslant S$. Our goal is to recover α as accurately as possible.

In the setting considered here we are allowed to select the points $\{x_n\}_{1 \leqslant n \leqslant N} \in \mathcal{X}$ where the function f is evaluated, which results in the noisy observations

$$y_n = f(x_n) + \eta_n, \tag{7.1}$$

where η_n is an observation noise term. We further assume that the noise is bounded, i.e. that $\|\eta\|_2^2 \stackrel{\text{def}}{=} \sum_{n=1}^{N} \eta_n^2 \leqslant \sigma^2$. We write $\mathcal{D}_N = (\{x_n, y_n\}_{1 \leqslant n \leqslant N})$ the set of observations and we are interested in situations where $N \ll K$, i.e., the number of observations is much smaller than the number of features φ_k.

The question we wish to address is: how well can we recover α based on a set of N noisy measurements? Note that whenever the noise is non-zero, the recovery cannot be perfect, so we wish to express the estimation error $\|\alpha - \widehat{\alpha}\|_2$ in terms of N, where $\widehat{\alpha}$ is our estimate.

The proposed method. We address the problem of sparse recovery by combining the two ideas:

- Sparse recovery theorems (see Section 2) essentially say that in order to recover a vector with a small number of measurements, one needs *incoherence*. The measurement basis, corresponding to the pointwise evaluations $f(x_n)$, should to be *incoherent* with the representation basis, corresponding to the one on which the vector α is sparse. Interpreting these basis in terms of linear operators, pointwise evaluation of f is equivalent to measuring f using Dirac masses $\delta_{x_n}(f) \stackrel{\text{def}}{=} f(x_n)$. Since in general the representation basis $\{\varphi_k\}_{1 \leqslant k \leqslant K}$ is not incoherent with the measurement basis induced by Dirac operators, we would like to consider another measurement basis, possibly randomized, in order that it becomes incoherent with any representation basis.

- Since we are interested in reconstructing α, and since we assumed that f is linear in α, we can apply any set of M linear operators $\{T_m\}_{1\leqslant m\leqslant M}$ to $f = \sum_k \alpha_k \varphi_k$, and consider the problem transformed by the operators; the parameter α is thus also the solution to the transformed problem $T_m(f) = \sum_k \alpha_k T_m(\varphi_k)$.

Thus, instead of considering the $N \times K$ sensing matrix $\Phi = (\delta_{x_n}(\varphi_k))_{k,n}$, we consider a new $M \times K$ sensing matrix $A = (T_m(\varphi_k))_{k,m}$, where the operators $\{T_m\}_{1\leqslant m\leqslant M}$ enforce incoherence between bases. Provided that we can estimate $T_m(f)$ with the data set $\mathcal{D}_N$, we will be able to recover α. The *Brownian sensing* approach followed here uses stochastic integral operators $\{T_m\}_{1\leqslant m\leqslant M}$, which makes the measurement basis incoherent with any representation basis, and generates a sensing matrix A which is Gaussian (with i.i.d. rows).

The proposed algorithm (detailed in Section 3) recovers α by solving the system $A\alpha \approx \widehat{b}$ by l_1 minimization[1], where $\widehat{b} \in \mathbb{R}^M$ is an estimate, based on the noisy observations y_n, of the vector $b \in \mathbb{R}^M$ whose components are $b_m = T_m f$.

Contribution: Our contribution is a sparse recovery result for *arbitrary non-orthonormal* functional basis $\{\varphi_k\}_{k\leqslant K}$ of a Hölder continuous function f. Theorem 7.4 states that our estimate $\widehat{\alpha}$ satisfies $\|\alpha - \widehat{\alpha}\|_2 = O(\|\eta\|_2/\sqrt{N})$ with high probability *whatever* N, under the assumption that the noise η is globally bounded, such as in Candés and Romberg (2007), Rauhut (2010). This result is obtained by combining two contributions:

- We show that when the sensing matrix A is Gaussian, i.e. when each row of the matrix is drawn i.i.d. from a Gaussian distribution, orthonormality is not required for sparse recovery. This result, stated in Proposition 7.1 (and used in Step 1 of the proof of Theorem 7.4), is a consequence of Theorem 3.1 of Foucart and Lai (2009).

- The sensing matrix A is made Gaussian by choosing the operators T_m to be stochastic integrals: $T_m f \stackrel{\text{def}}{=} \frac{1}{\sqrt{M}} \int_{\mathcal{C}} f dB^m$, where B^m are Brownian motions, and $\mathcal{C}$ is a 1-dimensional curve of $\mathcal{X}$ appropriately chosen according to the functions $\{\varphi_k\}_{k\leqslant K}$ (see the discussion in Section 4). We call A the *Brownian sensing* matrix.

We have the property that the recovery property using the Brownian sensing matrix A only depends on the number of Brownian motions M used in the stochastic integrals and not on the number of sampled points N. Note that M can be chosen arbitrarily large as it is not linked with the limited amount of data, but M affects the overall computational complexity of the method. The number of sample N appears in the quality of estimation of b only, and this is where the assumption that f is Hölder continuous comes into the picture.

Outline: In Section 2, we survey the large body of existing results about sparse recovery and relate our contribution to this literature. In Section 3, we explain in detail the Brownian sensing recovery method sketched above and state our main result in Theorem 7.4.

In Section 4, we first discuss our result and compare it with existing work. Then we comment on the choice and influence of the sampling domain $\mathcal{C}$ on the recovery performance.

[1] where the approximation sign $\approx$ refers to a minimization problem under a constraint coming from the observation noise.

Finally in Section 5, we report numerical experiments illustrating the recovery properties of the Brownian sensing method, and the benefit of the latter compared to a straightforward application of compressed sensing when there is noise and very few sampling points.

2 Relation to existing results

A standard approach in order to recover α[2] is to consider the $N \times K$ matrix $\Phi = (\varphi_k(x_n))_{k,n}$, and solve the system $\Phi\widehat{\alpha} \approx y$ where y is the vector with components y_n. Since $N \ll K$ this is an ill-posed problem. Under the sparsity assumption, a successful idea is first to replace the initial problem with the well-defined problem of minimizing the ℓ_0 norm of α under the constraint that $\Phi\widehat{\alpha} \approx y$, and then, since this problem is NP-hard, use convex relaxation of the ℓ_0 norm by replacing it with the ℓ_1 norm. We then need to ensure that the relaxation provides the same solution as the initial problem making use of the ℓ_0 norm. The literature on this problem is huge (see Candés and Romberg (2007), Donoho (2006), Donoho and Stark (1989), Tibshirani (1994), Zhao and Yu (2006), Candés and Tao (2007), Koltchinskii (2009) for examples of papers that initiated this field of research).

Generally, we can decompose the reconstruction problem into two distinct sub-problems. The first sub-problem (a) is to state conditions on the matrix Φ ensuring that the recovery is possible and derive results for the estimation error under such conditions:

The first important condition is the *Restricted Isometry Property* (RIP), introduced in Candés et al. (2006a), from which we can derive the following recovery result stated in Candés et al. (2006b):

Theorem 7.1 (Candés & al, 2006) *Let δ_S be the restricted isometry constant of $\frac{\Phi}{\sqrt{N}}$, defined as $\delta_S = \sup\{|\frac{\|\frac{\Phi}{\sqrt{N}}a\|_2}{\|a\|_2} - 1|; \|a\|_0 \leqslant S\}$. Then if $\delta_{3S} + \delta_{4S} < 2$, for every S-sparse vector $\alpha \in \mathbb{R}^K$, the solution $\widehat{\alpha}$ to the ℓ_1-minimization problem $\min\{\|a\|_1; a \text{ satisfies } \|\Phi a - y\|_2^2 \leqslant \sigma^2\}$ satisfies*

$$\|\widehat{\alpha} - \alpha\|_2^2 \leqslant \frac{C_S \sigma^2}{N},$$

where C_S depends only on δ_{4S}.

Additional recent results on that property are to be found in Candès (2008).

Apart from the historical RIP, many other conditions emerged from works reporting the practical difficulty to have the RIP satisfied, and thus weaker conditions ensuring reconstruction were derived. See van de Geer and Buhlmann (2009) for a precise survey of such conditions. A weaker condition for recovery is the *compatibility condition* which leads to the following result from van de Geer (2007):

Theorem 7.2 (Van de Geer & Buhlmann, 2009) *Assuming that the compatibility condition is satisfied, i.e. for a set $\mathcal{S}$ of indices of cardinality S and a constant L,*

[2]Note that reconstructing α is a more challenging and different goal than having a good approximation of the function f itself, as studied e.g. in Dalalyan and Tsybakov (2009).

$$C(L,\mathcal{S}) = \min\left\{\frac{S\|\frac{\Phi}{\sqrt{N}}\alpha\|_2^2}{\|\alpha_{\mathcal{S}}\|_1^2}, \alpha \text{ satisfies } \|\alpha_{\mathcal{S}^c}\|_1 \leqslant L\|\alpha_{\mathcal{S}}\|_1\right\} > 0,$$

then for every S-sparse vector $\alpha \in \mathbb{R}^K$, the solution $\widehat{\alpha}$ to the ℓ_1-minimization problem $\min\{\|\alpha\|_1; \alpha \text{ satisfies } \|\alpha_{\mathcal{S}^c}\|_1 \leqslant L\|\alpha_{\mathcal{S}}\|_1\}$ *satisfies for C a numerical constant:*

$$\|\widehat{\alpha} - \alpha\|_2^2 \leqslant \frac{C}{C(L,\mathcal{S})^2}\frac{\sigma^2\log(K)}{N}.$$

The second sub-problem (b) of the global reconstruction problem is to provide the user with a simple way to efficiently sample the space in order to build a matrix Φ such that the conditions for recovery are fulfilled, at least with high probability. This can be difficult in practice since it involves understanding the geometry of high dimensional objects. For instance, to the best of our knowledge, there is no result explaining how to sample the space so that the corresponding sensing matrix Φ satisfies the nice recovery properties needed by the previous theorems, for a *general* family of features $\{\varphi_k\}_{k\leqslant K}$.

However, it is proven in Rauhut (2010) that under some hypotheses on the functional basis, we are able to recover the strong RIP property for the matrix Φ with high probability. This result, combined with a recovery result, is stated as follows:

Theorem 7.3 (Rauhut, 2010) *Assume that $\{\varphi_k\}_{k\leqslant K}$ is an orthonormal basis of functions under a measure ν, bounded by a constant C_φ, and that we build $\mathcal{D}_N$ by sampling f at random according to ν. Assume also that the noise is bounded $\|\eta\|_2 \leqslant \sigma$. If $\frac{N}{\log(N)} \geqslant c_0 C_\varphi^2 S\log(S)^2\log(K)$ and $N \geqslant c_1 C_\varphi^2 S\log(p^{-1})$, then with probability at least $1-p$, for every S-sparse vector $\alpha \in \mathbb{R}^K$, the solution $\widehat{\alpha}$ to the ℓ_1-minimization problem* $\min\{\|a\|_1; a \text{ satisfies } \|Aa - y\|_2^2 \leqslant \sigma^2\}$ *satisfies*

$$\|\widehat{\alpha} - \alpha\|_2^2 \leqslant \frac{c_2\sigma^2}{N},$$

where c_0, c_1 and c_2 are some numerical constants.

In order to prove this theorem, the author of Rauhut (2010) showed that by sampling the points i.i.d. from ν, then with *with high probability* the resulting matrix Φ is RIP. The strong point of this Theorem is that we do not need to check conditions *on the matrix* Φ to guarantee that it is RIP, which is in practice infeasible. But the weakness of the result is that the initial basis has to be *orthonormal* and *bounded* under the given measure ν in order to get the RIP satisfied: the two conditions ensure incoherence with Dirac observation basis. The *specific* case of an unbounded basis i.e., Legendre Polynomial basis, has been considered in Rauhut and Ward (2010), but to the best of our knowledge, the problem of designing a *general* sampling strategy such that the resulting sensing matrix possesses nice recovery properties in the case of *non-orthonormal basis* remains unaddressed. Our contribution considers this case and is described in the following section.

3 The "Brownian sensing" approach

A need for incoherence. When the representation and observation basis are not incoherent, the sensing matrix Φ does not possess a nice recovery property. A natural idea is to change the observation basis by introducing a set of M *linear operators* $\{T_m\}_{m\leqslant M}$ acting on the functions $\{\varphi_k\}_{k\leqslant K}$. We have $T_m(f) = \sum_{k=1}^{K} \alpha_k T_m(\varphi_k)$ for all $1 \leqslant m \leqslant M$ and our goal is to define the operators $\{T_m\}_{m\leqslant M}$ in order that the sensing matrix $(T_m(\varphi_k))_{m,k}$ enjoys a nice recovery property, whatever the representation basis $\{\varphi_k\}_{k\leqslant K}$.

The Brownian sensing operators. We now consider linear operators defined by stochastic integrals on a 1-dimensional curve $\mathcal{C}$ of $\mathcal{X}$. First, we need to select a curve $\mathcal{C} \subset \mathcal{X}$ of length l, such that the covariance matrix $V_{\mathcal{C}}$, defined by its elements $(V_{\mathcal{C}})_{i,j} = \int_{\mathcal{C}} \varphi_i \varphi_j$ (for $1 \leqslant i, j \leqslant K$), is invertible. We will discuss the existence of a such a curve later in Section 4. Then, we define the linear operators $\{T_m\}_{1\leqslant m\leqslant M}$ as stochastic integrals over the curve $\mathcal{C}$: $T_m(g) \stackrel{\text{def}}{=} \frac{1}{\sqrt{M}} \int_{\mathcal{C}} g dB^m$, where $\{B^m\}_{m\leqslant M}$ are M independent Brownian motions defined on $\mathcal{C}$.

Note that up to an appropriate speed-preserving parametrization $g : [0, l] \to \mathcal{X}$ of $\mathcal{C}$, we can work with the corresponding induced family $\{\psi_k\}_{k\leqslant K}$, where $\psi_k = \varphi_k \circ g$, instead of the family $\{\varphi_k\}_{k\leqslant K}$.

The sensing method. With the choice of the linear operators $\{T_m\}_{m\leqslant M}$ defined above, the parameter $\alpha \in \mathbb{R}^K$ now satisfies the following equation

$$A\alpha = b\,, \tag{7.2}$$

where $b \in \mathbb{R}^M$ is defined by its components $b_m \stackrel{\text{def}}{=} T_m(f) = \frac{1}{\sqrt{M}} \int_{\mathcal{C}} f(x) dB^m(x)$ and the so-called Brownian sensing matrix A (of size $M \times K$) has elements $A_{m,k} \stackrel{\text{def}}{=} T_m(\varphi_k)$. Note that we do not require sampling f in order to compute the elements of A. Thus, the samples only serve for estimating b and for this purpose, we sample f at points $\{x_n\}_{1\leqslant n\leqslant N}$ regularly chosen along the curve $\mathcal{C}$.

In general, for a curve $\mathcal{C}$ parametrized with speed-preserving parametrization $g : [0, l] \to \mathcal{X}$ of $\mathcal{C}$, we have $x_n = g(\frac{n}{N} l)$ and the resulting estimate $\widehat{b} \in \mathbb{R}^M$ of b is defined with components:

$$\widehat{b}_m = \frac{1}{\sqrt{M}} \sum_{n=0}^{N-1} y_n (B^m(x_{n+1}) - B^m(x_n))\,. \tag{7.3}$$

Note that in the special case when $\mathcal{X} = \mathcal{C} = [0, 1]$, we simply have $x_n = \frac{n}{N}$.

The final step of the proposed method is to apply standard recovery techniques (e.g., l_1 minimization or Lasso) to compute $\widehat{\alpha}$ for the system (7.2) where b is perturbed by the so-called sensing noise $\varepsilon \stackrel{\text{def}}{=} b - \widehat{b}$ (estimation error of the stochastic integrals).

3.1 Properties of the transformed objects

We now give two properties of the Brownian sensing matrix A and the sensing noise $\varepsilon = b - \widehat{b}$.

Brownian sensing matrix. By definition of the stochastic integral operators $\{T_m\}_{m \leqslant M}$, the sensing matrix $A = (T_m(\varphi_k))_{m,k}$ is a centered Gaussian matrix, with

$$\mathrm{Cov}(A_{m,k}, A_{m,k'}) = \frac{1}{M} \int_{\mathcal{C}} \varphi_k(x) \varphi_{k'}(x) dx \,.$$

Moreover by independence of the Brownian motions, each row $A_{m,\cdot}$ is i.i.d. from a centered Gaussian distribution $N(0, \frac{1}{M} V_{\mathcal{C}})$, where $V_{\mathcal{C}}$ is the $K \times K$ covariance matrix of the basis, defined by its elements $V_{k,k'} = \int_{\mathcal{C}} \varphi_k(x) \varphi_{k'}(x) dx$.

Thanks to this nice structure, we can prove that A possesses a property similar to RIP (in the sense of Foucart and Lai (2009)) whenever M is large enough:

Proposition 7.1 *For $p > 0$ and any integer $t > 0$, when $M > \frac{C'}{4}(t \log(K/t) + \log 1/p))$, with C' being a universal constant (defined in Rudelson and Vershynin (2008b), Baraniuk et al. (2008)), then with probability at least $1 - p$, for all $t-$sparse vectors $x \in \mathbb{R}^K$,*

$$\frac{1}{2} \nu_{\min,\mathcal{C}} \|x\|_2 \leqslant \|Ax\|_2 \leqslant \frac{3}{2} \nu_{\max,\mathcal{C}} \|x\|_2,$$

where $\nu_{\max,\mathcal{C}}$ and $\nu_{\min,\mathcal{C}}$ are respectively the largest and smallest eigenvalues of $V_{\mathcal{C}}^{1/2}$.

Sensing noise. In order to state our main result, we need a bound on $\|\varepsilon\|_2^2$. We consider the simplest deterministic *sensing design* where we choose the sensing points to be uniformly distributed along the curve $\mathcal{C}$[3].

Proposition 7.2 *Assume that $\|\eta\|_2^2 \leqslant \sigma^2$ and that f is (L, β)-Hölder, i.e.*

$$\forall (x, y) \in \mathcal{X}^2, |f(x) - f(y)| \leqslant L|x - y|^\beta \,,$$

then for any $p \in (0, 1]$, with probability at least $1 - p$, we have the following bound on the sensing noise $\varepsilon = b - \widehat{b}$ computed on the curve $\mathcal{C}$ of length l:

$$\|\varepsilon\|_2^2 \leqslant \frac{\tilde{\sigma}^2(N, M, p)}{N} \,,$$

where

$$\tilde{\sigma}^2(N, M, p) \stackrel{\text{def}}{=} 2\Big(\frac{L^2 l^{2\beta}}{N^{2\beta-1}} + \sigma^2\Big)\Big(1 + 2\frac{\log(1/p)}{M} + 4\sqrt{\frac{\log(1/p)}{M}}\Big) \,.$$

Remark 5 *The bound on the sensing noise $\|\varepsilon\|_2^2$ contains two contributions: an approximation error term which comes from the approximation of a stochastic integral with N points and that scales with $L^2 l^{2\beta}/N^{2\beta}$, and the observation noise term of order σ^2/N. The observation noise term (when $\sigma^2 > 0$) dominates the approximation error term whenever $\beta \geqslant 1/2$.*

Input: a curve $\mathcal{C}$ of length l such that $V_{\mathcal{C}}$ is invertible. Parameters N and M.

- Select N uniform samples $\{x_n\}_{1\leqslant n\leqslant N}$ along the curve $\mathcal{C}$,
- Generate M Brownian motions $\{B^m\}_{1\leqslant m\leqslant M}$ along $\mathcal{C}$.
- Compute the Brownian sensing matrix $A \in \mathbb{R}^{M\times K}$
$$\text{(i.e. } A_{m,k} = \tfrac{1}{\sqrt{M}} \textstyle\int_{\mathcal{C}} \varphi_k(x) dB^m(x)\text{)}.$$
- Compute the estimate $\widehat{b} \in \mathbb{R}^M$
$$\text{(i.e. } \widehat{b}_m = \tfrac{1}{\sqrt{M}} \textstyle\sum_{n=0}^{N-1} y_n (B^m(x_{n+1}) - B^m(x_n))\text{)}.$$
- Find $\widehat{\alpha}$, solution to
$$\min_a \left\{ \|a\|_1 \;\text{ such that }\; \|Aa - \widehat{b}\|_2^2 \leqslant \frac{\tilde{\sigma}^2(N,M,p)}{N} \right\}.$$

Figure 7.1: The Brownian sensing approach using a uniform sampling along the curve $\mathcal{C}$.

3.2 Main result.

In this section, we state our main recovery result for the Brownian sensing method, described in Figure 7.1, using a uniform sampling method along a one-dimensional curve $\mathcal{C} \subset \mathcal{X} \subset \mathbb{R}^d$. The proof of the following theorem can be found in the supplementary material.

Theorem 7.4 (Main result) *Assume that f is (L,β)-Hölder on $\mathcal{X}$ and that $V_{\mathcal{C}}$ is invertible. Let us write the condition number $\kappa_{\mathcal{C}} = \nu_{\max,\mathcal{C}}/\nu_{\min,\mathcal{C}}$, where $\nu_{\max,\mathcal{C}}$ and $\nu_{\min,\mathcal{C}}$ are respectively the largest and smallest eigenvalues of $V_{\mathcal{C}}^{1/2}$. Write $r = \left[(3\kappa_{\mathcal{C}} - 1)(\frac{1}{4\sqrt{2}-1})\right]^2$. For any $p \in (0,1]$, let $M \geqslant 4c(4Sr\log(\frac{K}{4Sr}) + \log 1/p)$ (where c is a universal constant defined in Rudelson and Vershynin (2008b), Baraniuk et al. (2008)). Then, with probability at least $1-3p$, the solution $\widehat{\alpha}$ obtained by the Brownian sensing approach described in Figure 7.1, satisfies*

$$\|\widehat{\alpha} - \alpha\|_2^2 \leqslant C\Big(\frac{\kappa_{\mathcal{C}}^4}{\max_k \int_{\mathcal{C}} \varphi_k^2}\Big)\frac{\tilde{\sigma}^2(N,M,p)}{N},$$

where C is a numerical constant and $\tilde{\sigma}(N,M,p)$ is defined in Proposition 7.2.

Note that a similar result (not reported here) can be proven in the case of i.i.d. sub-Gaussian noise, instead of a noise with bounded ℓ_2 norm considered here.

[3] Note that other deterministic, random, or low-discrepancy sequence could be used here.

4 Discussion.

In this section we discuss the differences with previous results, especially with the work Rauhut (2010) recalled in Theorem 7.3. We then comment on the choice of the curve C and illustrate examples of such curves for different bases.

4.1 Comparison with known results

The order of the bound. Concerning the scaling of the estimation error in terms of the number of sensing points N, Theorem 7.3 of Rauhut (2010) (reminded in Section 2) states that when N is large enough (i.e., $N = \Omega(S\log(K))$), we can build an estimate $\widehat{\alpha}$ such that $\|\widehat{\alpha} - \alpha\|_2^2 = O(\frac{\sigma^2}{N})$. In comparison, our bound shows that $\|\widehat{\alpha} - \alpha\|_2^2 = O(\frac{L^2 l^{2\beta}}{N^{2\beta}} + \frac{\sigma^2}{N})$ *for any values* of N. Thus, provided that the function f has a Hölder exponent $\beta \geqslant 1/2$, we obtain the same rate as in Theorem 7.3.

A weak assumption about the basis. Note that our recovery performance scales with the condition number $\kappa_{\mathcal{C}}$ of $V_{\mathcal{C}}$ as well as the length l of the curve $\mathcal{C}$. However, concerning the hypothesis on the functions $\{\varphi_k\}_{k\leqslant K}$, we only assume that the covariance matrix $V_{\mathcal{C}}$ is invertible on the curve $\mathcal{C}$, which enables to handle *arbitrarily non-orthonormal bases*. This means that the orthogonality condition on the basis functions is not a crucial requirement to deduce sparse recovery properties. To the best of our knowledge, this is an improvement over previously known results (such as the work of Rauhut (2010)). Note however that if $\kappa_{\mathcal{C}}$ or l are too high, then the bound becomes loose. Also the computational complexity of the Brownian sensing increases when $\kappa_{\mathcal{C}}$ is large, since it is necessary to take a large M, i.e. to simulate more Brownian motions in that case.

A result that holds without any conditions on the number of sampling points. Theorem 7.4 requires a constraint on the number of Brownian motions M (i.e., that $M = \Omega(S\log K)$) and not on the number of sampling points N (as in Rauhut (2010), see Theorem 7.3). This is interesting in practical situations when we do not know the value of S, as we do not have to assume a lower-bound on N to deduce the estimation error result. This is due to the fact that the Brownian sensing matrix A only depends on the computation of the M stochastic integrals of the K functions φ_k, and does not depend on the samples. The bound shows that we should take M as large as possible. However, M impacts the numerical cost of the method. This implies in practice a trade-off between a large M for a good estimation of α and a low M for low numerical cost.

4.2 The choice of the curve

Why sampling along a 1-dimensional curve $\mathcal{C}$ instead of sampling over the whole space $\mathcal{X}$? In a bounded space $\mathcal{X}$ of dimension 1, both approaches are identical. But

in dimension $d > 1$, following the Brownian sensing approach while sampling over the whole space would require generating M Brownian sheets (extension of Brownian motions to $d > 1$ dimensions) over $\mathcal{X}$, and then building the $M \times K$ matrix A with elements $A_{m,k} = \int_{\mathcal{X}} \varphi_k(t_1, ...t_d) dB_1^m(t_1)...dB_d^m(t_d)$. Assuming that the covariance matrix $V_{\mathcal{X}}$ is invertible, this Brownian sensing matrix is also Gaussian and enjoys the same recovery properties as in the one-dimensional case. However, in this case, estimating the stochastic integrals $b_m = \int_{\mathcal{X}} f dB^m$ using sensing points along a (d-dimensional) grid would provide an estimation error $\varepsilon = b - \widehat{b}$ that scales poorly with d since we integrate over a d dimensional space. This explains our choice of selecting a 1-dimensional curve $\mathcal{C}$ instead of the whole space $\mathcal{X}$ and sampling N points along the curve. This choice provides indeed a better estimation of b which is defined by a 1-dimensional stochastic integrals over $\mathcal{C}$. Note that the only requirement for the choice of the curve $\mathcal{C}$ is that the covariance matrix $V_{\mathcal{C}}$ defined along this curve should be invertible.

In addition, in some specific applications the sampling process can be very constrained by physical systems and sampling uniformly in all the domain is typically costly. For example in some medical experiments, e.g., scanner or I.R.M., it is only possible to sample along straight lines.

What the parameters of the curve tell us on a basis. In the result of Theorem 7.4, the length l of the curve $\mathcal{C}$ as well as the condition number $\kappa_{\mathcal{C}} = \nu_{\max,\mathcal{C}}/\nu_{\min,\mathcal{C}}$ are essential characteristics of the efficiency of the method. It is important to note that those two variables are actually related. Indeed, it may not be possible to find a short curve $\mathcal{C}$ such that $\kappa_{\mathcal{C}}$ is small. For instance in the case where the basis functions have compact support, if the curve $\mathcal{C}$ does not pass through the support of all functions, $V_{\mathcal{C}}$ will not be invertible. Any function whose support does not intersect with the curve would indeed be an eigenvector of $V_{\mathcal{C}}$ with a 0 eigenvalue. This indicates that the method will not work well in the case of a very localized basis $\{\varphi_k\}_{k \leqslant K}$ (e.g. wavelets with compact support), since the curve would have to cover the whole domain and thus l will be very large. On the other hand, the situation may be much nicer when the basis is not localized, as in the case of a Fourier basis. We show in the next subsection that in a d-dimensional Fourier basis, it is possible to find a curve $\mathcal{C}$ (actually a segment) such that the basis is orthonormal along the chosen line (i.e. $\kappa_{\mathcal{C}} = 1$).

4.3 Examples of curves

For illustration, we exhibit three cases for which one can easily derive a curve $\mathcal{C}$ such that $V_{\mathcal{C}}$ is invertible. The method described in the previous section will work with the following examples.

$\mathcal{X}$ is a segment of $\mathbb{R}$: In this case, we simply take $\mathcal{C} = \mathcal{X}$, and the sparse recovery is possible whenever the functions $\{\varphi_k\}_{k \leqslant K}$ are linearly independent in $\mathcal{L}_2$.

Coordinate functions: Consider the case when the basis are the coordinate functions $\varphi_k(t_1, ...t_d) = t_k$. Then we can define the parametrization of the curve $\mathcal{C}$ by setting $g(t) = \alpha(t)(t, t^2, \dots, t^d)$, where $\alpha(t)$ is the solution to a differential equation such that $\|g'(t)\|_2 = 1$ (which implies that for any function h, $\int h \circ g = \int_{\mathcal{C}} h$). The corresponding functions $\psi_k(t) = \alpha(t)t^k$ are linearly independent, since the only functions $\alpha(t)$ such that the $\{\psi_k\}_{k \leqslant K}$ are not linearly independent are functions that are 0 almost everywhere, which would contradict the definition of $\alpha(t)$. Thus $V_{\mathcal{C}}$ is invertible.

Fourier basis: Let us now consider the Fourier basis in $\mathbb{R}^d$ with frequency T:

$$\varphi_{n_1,\dots,n_d}(t_1, .., t_d) = \prod_j \exp\big(-\frac{2i\pi n_j t_j}{T}\big),$$

where $n_j \in \{0, ..., T-1\}$ and $t_j \in [0, 1]$. Note that this basis is orthonormal under the uniform distribution on $[0, 1]^d$. In this case we define g by $g(t) = \lambda(t\frac{1}{T^{d-1}}, t\frac{T}{T^{d-1}}, ..., t\frac{T^{d-1}}{T^{d-1}})$ with $\lambda = \sqrt{\frac{1-T^{-2}}{1-T^{-2d}}}$ (so that $\|g'(t)\|_2 = 1$), thus we deduce that:

$$\psi_{n_1,\dots,n_d}(t) = \exp\big(-\frac{2i\pi t\lambda \sum_j n_j T^{j-1}}{T^d}\big).$$

Since $n_k \in \{0, ..., T-1\}$, the mapping that associates $\sum_j n_j T^{j-1}$ to $(n_1, \dots, n_d)$ is a bijection from $\{0, \dots, T-1\}^d$ to $\{0, \dots, T^d - 1\}$. Thus we can identify the family $(\psi_{n_1,\dots,n_d})$ with the one dimensional Fourier basis with frequency $\frac{T^d}{\lambda}$, which means that the condition number $\rho = 1$ for this curve. Therefore, for a d-dimensional function f, sparse in the Fourier basis, it is sufficient to sample along the curve induced by g to ensure that $V_{\mathcal{C}}$ is invertible.

5 Numerical Experiments

In this section, we illustrate the method of Brownian sensing in dimension one. We consider a non-orthonormal family $\{\varphi_k\}_{k \leqslant K}$ of $K = 100$ functions of $L_2([0, 2\pi])$ defined by $\varphi_k(t) = \frac{\cos(tk)+\cos(t(k+1))}{\sqrt{2}\pi}$. In the experiments, we use a function f whose decomposition is 3-sparse and which is $(10, 1)$-Hölder, and we consider a bounded observation noise η, with different noise levels, where the noise level is defined by $\sigma^2 = \sum_{n=1}^N \eta_n^2$.

In Figure 7.2, the plain curve represents the recovery performance, i.e., the mean squared error of Brownian sensing that is minimizing $\|a\|_1$ under the constraint that $\|Aa - \widehat{b}\|_2 \leqslant 1.95\sqrt{2(100/N + 2)}$ using $M = 100$ Brownian motions and a regular grid of N points, as a function of N[4]. The dashed curve represents the mean squared error of a regular l_1 minimization of $\|a\|_1$ under the constraint that $\|\Phi a - y\|_2^2 \leqslant \sigma^2$ (as described e.g. in Rauhut (2010)), where the N samples are drawn uniformly randomly over the domain. The three different graphics correspond to different values of the noise level σ^2 (from left to right 0, 0.5 and 1). Note that the results are averaged over 5000 trials.

[4]We assume that we know a loose bound on the noise level, here $\sigma^2 \leqslant 2$, and we take $p = 0.01$.

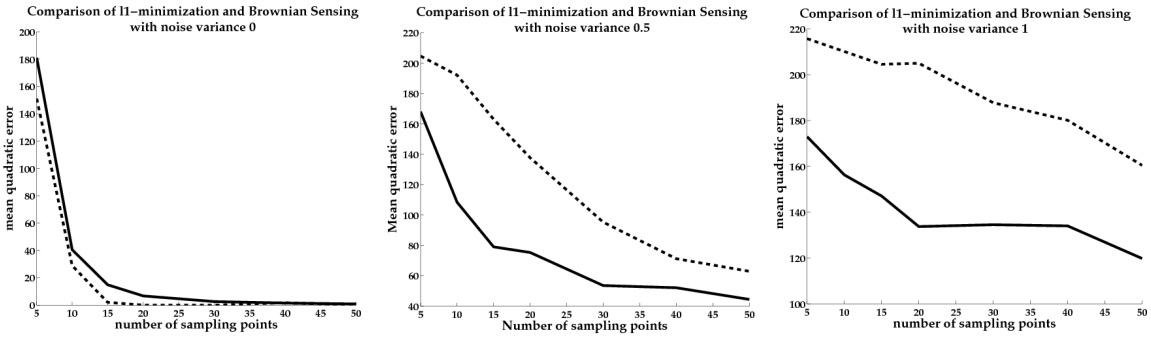


Figure 7.2: Mean squared estimation error using Brownian sensing (plain curve) and a direct l_1-minimization solving $\Phi\alpha \approx y$ (dashed line), for different noise level ($\sigma^2 = 0$, $\sigma^2 = 0.5$, $\sigma^2 = 1$), plotted as a function of the number of sample points N.

Figure 7.2 illustrates that, as expected, Brownian sensing outperforms the method described in Rauhut (2010) for noisy measurements[5]. Note also that the method described in Rauhut (2010) recovers the sparse vector when there is no noise, and that Brownian sensing in this case has a smoother dependency w.r.t. N. Note that this improvement comes from the fact that we use the Hölder regularity of the function: Compressed sensing may outperform Brownian sensing for arbitrarily non regular functions.

Conclusion

In this chapter, we have introduced a so-called Brownian sensing approach, as a way to sample an unknown function which has a sparse representation on a given non-orthonormal basis. Our approach differs from previous attempts to apply compressed sensing in the fact that we build a "Brownian sensing" matrix A based on a set of Brownian motions, which is independent of the function f. This enables us to guarantee nice recovery properties of A. The function evaluations are used to estimate the right hand side term b (stochastic integrals). In dimension d we proposed to sample the function along a well-chosen curve, i.e. such that the corresponding covariance matrix is invertible. We provided competitive reconstruction error rates of order $O(\|\eta\|_2/\sqrt{N})$ when the observation noise η is bounded and f is assumed to be Hölder continuous with exponent at least $1/2$. We believe that the Hölder assumption is not strictly required (the smoothness of f is assumed to derive nice estimations of the stochastic integrals only), and future works will consider weakening this assumption, possibly by considering randomized sampling designs.

[5]Note however that there is no theoretical guarantee that the method described in Rauhut (2010) works here since the functions are not orthonormal.

6 Technical details

Proof of Proposition 7.1

First, we prove a very short Lemma describing some properties of the matrix A.

Lemma 7.1 *Let us consider M independent Brownian motions $(B^1, ..., B^M)$ on $\mathcal{X}$, and define the $M \times K$ matrix A with elements*

$$A_{m,k} = \frac{1}{\sqrt{M}}\Big(\int_{\mathcal{C}} \varphi_k(x) dB^m(x)\Big).$$

Then A is a centered Gaussian matrix where each row $A_{m,\cdot}$ is i.i.d. from $\mathcal{N}(0, \frac{1}{M}V_{\mathcal{C}})$, where $V_{\mathcal{C}}$ is the $K \times K$ covariance matrix of the basis, defined by its elements $V_{k,k'} = \int_{\mathcal{C}} \varphi_k(x)\varphi_{k'}(x)dx$.

Proof: Indeed, from the definition of stochastic integrals, each $A_{m,k} \sim \mathcal{N}(0, \frac{1}{M}\int_{\mathcal{C}} \varphi_k^2(x)dx)$, and $\mathrm{Cov}(A_{m,k}, A_{m,k'}) = \frac{1}{M}\int_{\mathcal{C}} \varphi_k(x)\varphi_{k'}(x)dx$. Thus each row $A_{m,\cdot} \sim \mathcal{N}(0, \frac{1}{M}V_{\mathcal{C}})$ and are independent by independence of the Brownian motions. Additionally, we have

$$\mathbb{E}[(A^TA)_{k,k'}] = \mathbb{E}\Big[\frac{1}{M}\sum_{m=1}^{M} A_{m,k}A_{m,k'}\Big] = V_{k,k',\mathcal{C}}.$$

□

Now let us define $B = AV_{\mathcal{C}}^{-1/2}$. Since each row of A is an independent draw of $\mathcal{N}(0, V_{\mathcal{C}})$, then each row of B is an independent draw of $\mathcal{N}(0, I)$. Thus B is a matrix with elements i.i.d. from $\mathcal{N}(0, 1)$. We thus can use the following result (as stated in Fornasier and Rauhut (to appear), see also Rudelson and Vershynin (2008b), Baraniuk et al. (2008)):

Theorem 7.5 *For $p' > 0$ and any integer $t > 0$, when $M > C'\delta^{-2}(t\log(K/t) + \log 1/p'))$, with C' being a universal constant, see Rudelson and Vershynin (2008b), Baraniuk et al. (2008), then with probability at least $1 - p'$, there exists $\delta_t \leqslant \delta$ (δ_t is the RIP constant of B for t-sparse vectors) such that for all $t-$sparse vectors $x \in \mathbb{R}^K$,*

$$(1 - \delta_t)\|x\|_2 \leqslant \|Bx\|_2 \leqslant (1 + \delta_t)\|x\|_2.$$

Since $V_{\mathcal{C}}$ is symmetric, it is possible to write $V_{\mathcal{C}} = UDU^T$ with U an orthogonal matrix and D a diagonal matrix with the eigenvalues of V as diagonal elements (SVD decomposition). Thus, $V^{1/2} = UD^{1/2}U^T$ where $D^{1/2}$ is the diagonal matrix with the square roots of the diagonal elements of D (i.e., the eigenvalues of $V_{\mathcal{C}}^{1/2}$).

Note that if U is an orthogonal matrix, BU is also RIP with the same constant as B (see Donoho (2006) for the preservation of the RIP property to a change of orthonormal basis). Applying this and Theorem 7.5 with $\delta = 1/2$ for $2t$-sparse vectors, we have that

whenever $M > 4C'(2t\log(K/2t) + \log 1/p')$, the RIP constant δ_{2t} satisfies $\delta_{2t} \leqslant 1/2$, i.e. for all $2t-$sparse vectors x,

$$\frac{1}{2}\|x\|_2 \leqslant \|BUx\|_2 \leqslant \frac{3}{2}\|x\|_2.$$

Now if we consider a $2t-$sparse vector x, then $D^{1/2}x$ is also $2t-$sparse with same support as x, and we also have that $\nu_{min,\mathcal{C}}\|x\|_2 \leqslant \|D^{1/2}x\|_2 \leqslant \nu_{max,\mathcal{C}}\|x\|_2$. Thus the matrix $BUD^{1/2}$ satisfies

$$\frac{\nu_{min,\mathcal{C}}}{2}\|x\|_2 \leqslant \|BUD^{1/2}x\|_2 \leqslant \frac{3\nu_{max,\mathcal{C}}}{2}\|x\|_2.$$

As mentioned before, the preservation of the RIP property to a change of orthonormal base (see Donoho (2006)) can be applied with U and thus as $A = BV^{1/2} = BUD^{1/2}U^T$ to obtain:

$$\frac{1}{2}\nu_{\min,\mathcal{C}}\|x\|_2 \leqslant \|Ax\|_2 \leqslant \frac{3}{2}\nu_{\max,\mathcal{C}}\|x\|_2.$$

Proof of Proposition 7.2

We prove here without loss of generality (because of we can always parametrize the curve) the result for $\mathcal{X} = [0, l]$. Let us recall that f is (L, β)-Hölder and that we write $\sigma = \|\eta\|_2$. The estimation error $\varepsilon_m = b_m - \widehat{b_m}$, given the samples $(x_n, y_n)_n$, follows a centered Gaussian distribution (w.r.t. the choice of the Brownian B^m) with variance

$$\begin{aligned}
\mathbb{V}(\varepsilon_m) &= \mathbb{V}\left(\frac{1}{\sqrt{M}}\Big(\int_0^l f(x)dB^m(x) - \sum_{n=0}^{N-1} y_n(B^m_{x_{n+1}} - B^m_{x_n})\Big)\right) \\
&= \frac{1}{M}\mathbb{V}\left(\int_0^l \big(f(x) - \sum_n (f(l\frac{(n+1)}{N}) + \eta_n)\mathbb{I}_{x\in[l\frac{n}{N};l\frac{(n+1)}{N}]}\big)dB^m(x)\right) \\
&= \frac{1}{M}\int_0^l \big(f(x) - \sum_n (f(l\frac{n}{N}) + \eta_n)\mathbb{I}_{x\in[l\frac{n}{N};l\frac{(n+1)}{N}]}\big)^2 dx \\
&= \frac{1}{M}\sum_n \int_{l\frac{n}{N}}^{l\frac{(n+1)}{N}} (f(x) - f(l\frac{n}{N}) - \eta_n)^2 dx \\
&\leqslant \frac{1}{MN}\sum_n (\frac{Ll^\beta}{N^\beta} + |\eta_n|)^2 dx \\
&= \frac{2}{MN}\Big(\frac{L^2 l^{2\beta}}{N^{2\beta-1}} + \sum_n |\eta_n|^2\Big) \\
&\leqslant \frac{2}{MN}\Big(\frac{L^2 l^{2\beta}}{N^{2\beta-1}} + \sigma^2\Big).
\end{aligned}$$

We now wish to apply Bernstein's inequality in order to bound $\|\varepsilon\|_2$ in high probability. We recall the following result (see e.g. Bennett (1962)):

Theorem 7.6 (Bernstein's inequality) *Let $(X_1, ...X_M)$ be independent real valued random variables and assume that there exist two positive numbers v and d such that: $\sum_{m=1}^M \mathbb{E}(X_m^2) \leqslant v$ and for all integers $r \geqslant 3$,*

$$\sum_{m=1}^M \mathbb{E}[(X_m)_+^r] \leqslant \frac{r!}{2} v d^{r-2}.$$

Let $S = \sum_{m=1}^M (X_m - \mathbb{E}(X_m))$, then for any $x \geqslant 0$, we have $\mathbb{P}(S \geqslant \sqrt{2vx} + dx) \leqslant \exp(-x)$.

Let us check that the assumptions for applying Bernstein's inequality hold with the choice $v = 8M(\mathbb{V}(\varepsilon_m))^2$ and $d = 2\mathbb{V}(\varepsilon_m)$. Indeed, since the ε_m are i.i.d. centered Gaussian, by writing $X_m = \varepsilon_m^2$, we have $X_m \geqslant 0$ and for any integer $r \geqslant 2$, $\mathbb{E}(X_m^r) = (\mathbb{V}(\varepsilon_m))^r \frac{(2r)!}{2^r r!}$. This gives $\sum_{m=1}^M \mathbb{E}[X_m^2] = 3M(\mathbb{V}(\varepsilon_m))^2 \leqslant v$, and for $r \geqslant 3$,

$$\sum_{m=1}^M \mathbb{E}[X_m^r] = M(\mathbb{V}(\varepsilon_m))^r \frac{(2r)!}{2^r r!} \leqslant M(\mathbb{V}(\varepsilon_m))^r \times 2^r r! \leqslant \frac{r!}{2} v d^{r-2}.$$

We thus apply Bernstein's inequality (and recall that $\mathbb{V}(\varepsilon_m) \leqslant \frac{2}{MN}\left(\frac{L^2 l^{2\beta}}{N^{2\beta-1}} + \sigma^2\right)$) to obtain that with probability at least $1-p$,

$$\|\varepsilon\|_2^2 \leqslant 2\Big(\frac{L^2 l^{2\beta}}{N^{2\beta}} + \frac{\sigma^2}{N}\Big)\Big(1 + 4\sqrt{\frac{\log(1/p)}{M}} + 2\frac{\log(1/p)}{M}\Big).$$

Proof of Theorem 7.4

Following Foucart and Lai (2009), we define $\alpha_t > 0$ (respectively $\beta_t > 0$) as the maximal (resp. minimal) values such that for all $x \in \mathbb{R}^K$ which are $t-$sparse,

$$\alpha_t \|x\|_2 \leqslant \|Ax\|_2 \leqslant \beta_t \|x\|_2. \tag{7.4}$$

We now define $\gamma_t = \frac{\beta_t}{\alpha_t}$ and use Theorem 3.1 of Foucart and Lai (2009) applied to sparse vectors, in the case of ℓ_1 minimization, reminded below:

Theorem 7.7 (Foucart, Lai) *For any integer $S > 0$, for $t \geqslant S$, whenever $\gamma_{2t} - 1 \leqslant 4(\sqrt{2}-1)\sqrt{\frac{t}{S}}$, the solution $\widehat{\alpha}$ to the ℓ_1-minimization problem*

$$\min \|a\|_1, \text{ under the constraint } \|Aa - b\|_2^2 \leqslant \|\varepsilon\|_2^2,$$

satisfies $\|\alpha - \widehat{\alpha}\|_2 \leqslant \frac{D_2 \|\varepsilon\|_2}{\beta_{2S}}$, where D_2 is a constant which depends on γ_{2t}, S and t defined in Foucart and Lai (2009).

In order to apply this results, we now provide conditions such that (7.4) holds, as well as an upper bound on the noise $\|\varepsilon^2\|$, and a lower bound on β_{2S}.

Step 1. Recovery Condition: We recall the results of Proposition 7.1 and have that (7.4) holds with $\alpha_{2t} \geqslant \frac{1}{2}\nu_{\min,\mathcal{C}}$ and $\beta_{2t} \leqslant \frac{3}{2}\nu_{\max,\mathcal{C}}$ with probability $1-p'$ as long as $M > \frac{C'}{4}(t\log(K/t) + \log 1/p'))$. Thus $\gamma_{2t} \leqslant 3\frac{\nu_{\max,\mathcal{C}}}{\nu_{\min,\mathcal{C}}} = 3\kappa_{\mathcal{C}}$.

A sufficient condition for (7.7) is that $3\kappa_{\mathcal{C}} - 1 \leqslant 4(\sqrt{2}-1)\sqrt{\frac{t}{S}}$.

By defining $r = \left[(3\kappa_{\mathcal{C}} - 1)(\frac{1}{4\sqrt{2}-1})\right]^2$ (note that r only depends on $V_{\mathcal{C}}$), condition (7.7) holds whenever $t > Sr$, thus with probability $1-p'$, whenever

$$M > 4C'\big(2\lceil Sr\rceil \log\frac{K}{2Sr} + \log 1/p'\big). \tag{7.5}$$

Note that this condition holds when the number of Brownian motions M (which can be chosen arbitrarily) is large enough (and does not depend on the number of observations N).

Step 2. Upper bound on $\|\varepsilon^2\|$: This is the result of Proposition 7.2.

Step 3. Lower bound on β_{2S} In order to apply Theorem 7.7, we now provide a lower bound on β_{2S}.

Lemma 7.2 *If*

$$M > C' \log 1/u, \tag{7.6}$$

then with probability $1-u$ we have: $\beta_{2S} \geqslant \frac{1}{2}\sqrt{\max_k \int_{\mathcal{C}} \varphi_k^2}$.

Proof: Let us define $i = \arg\max_k \int_{\mathcal{C}} \varphi_k^2(x)dx$. Let us now consider the $1-$sparse vector a such that $a_i = 1$ and $a_k = 0$ otherwise. We have: $(Aa)_m = \int_{\mathcal{C}} \varphi_i(x)dB^m(x)$. So each $(Aa)_m$ is a sample drawn independently from $\mathcal{N}(0, \int_{\mathcal{C}} \varphi_i^2(x)dx)$.

By applying Theorem 7.5, with $S = K = 1$ and $\delta = 1/2$, when $M > C' \log 1/u$, then with probability $1-u$,

$$\frac{1}{2}\sqrt{\int_{\mathcal{C}} \varphi_i^2(x)dx}\|a\|_2 \leqslant \|Aa\|_2 \leqslant \frac{3}{2}\sqrt{\int_{\mathcal{C}} \varphi_i^2(x)dx}\|a\|_2.$$

And since β_{2S} is the minimal constant such that for every $2S-$sparse vector x (in particular for a) we have $\|Ax\|_2 \leqslant \beta_{2S}\|x\|_2$, we deduce that

$$\beta_{2S} \geqslant \frac{1}{2}\sqrt{\int_{\mathcal{C}} \varphi_i^2(x)dx} = \frac{1}{2}\sqrt{\max_k \int_{\mathcal{C}} \varphi_k^2(x)dx}.$$

□

We now apply Theorem 7.7 and deduce that when M satisfies (7.5) (which implies that (7.6) also holds) using Lemma 7.2, with probability $1-p'-u$,

$$\|\widehat{\alpha} - \alpha\|_2 \leqslant \frac{2D_2\tilde{\sigma}(N,M,p)}{\sqrt{N}\sqrt{\max_k \int_{\mathcal{C}} \varphi_k^2}} \tag{7.7}$$

Thus from Proposition 7.2, with probability $1-p-p'-u$,

$$\|\widehat{\alpha}-\alpha\|_2^2 \leqslant \frac{8D_2^2\left(\frac{L^2}{N^{2\beta-1}}l^{2\beta}+\sigma^2\right)(1+c(p,M))}{N(\max_k \int_{\mathcal{C}}\varphi_k^2)},$$

and from Foucart and Lai (2009), we deduce that if we are able to recover $4S-$sparse vectors, i.e., if $M > 4C'\left(4Sr\log\frac{K}{4Sr}+\log 1/p'\right)$ then $D_2 \leqslant C\kappa_{\mathcal{C}}^2$ where C can be loosely bounded by 90, see Foucart and Lai (2009) (note that this numerical constant can be greatly improved). The result follows with the choice $p=p'=u$.

Chapter 8

Multiview Learning: Complexity versus Agreement.

In this chapter, we consider the problem of semi-supervised multiview classification, where we assume that each view corresponds to a Reproducing Kernel Hilbert Space. We study an algorithm based on co-regularization methods with extra penalty terms reflecting smoothness and general agreement properties. This work provides both an explicit upper and lower bound on the Rademacher complexity of the corresponding class of learners for an arbitrary large number of views. We also give asymptotic behavior of the bounds when the co-regularization term increases, making explicit the relation between *consistency* of the views and *reduction* of the search space. We apply this algorithm to some toy examples including a new challenging dataset. Finally, we advocate for a stability-based parameter selection inspired by clustering and localization arguments, give explicit bounds on the variance of the class and propose a selection algorithm.

This work was done while I was in the Master "Maths, Vision et Apprentissage" (MVA) of ENS-Cachan, and is joint work with *Nicolas Vayatis*. It has been published in the proceedings of the 20*th international conference on Algorithmic Learning Theory (ALT 2009)*, see Maillard and Vayatis (2009). Although this work is not fully connected with the rest of this Ph.D dissertation, note that the study of Rademacher complexities is very important in statistical learning in order to control the deviations of the empirical process indexed by a family of functions. Now in sequential learning, the corresponding notion has only been popularized very recently in a work that paves the way for important future research, see Rakhlin et al. (2010)). Thus understanding (local) Rademacher complexities and concepts around is clearly important for the development of sequential learning theory.

Contents

1 Introduction

We consider a classification problem in which each object to be classified may have many possible representations. For instance, movie classification can be based on the title only, on the analysis of the audio signal, or of some pattern analysis of some images... Now each distinct representation of an object will correspond to a different view (or representation space) for which a class of classifiers will be chosen, and we consider we have V many different views with $V \geq 2$, each of them begin a reproducing kernel Hilbert space (RKHS).

In the sequel, we assume that the learner knows how to represent objects in each representation space. This means that for each object, and for each representation space denoted by $\mathcal{X}_v$ for $v \in \{1, \ldots, V\}$, she can build a point $x^{(v)} \in \mathcal{X}_v$ representing this data. We write $x = (x^{(1)}, ..., x^{(V)})$ for the resulting point living in the product space $\prod_{v=1}^{V} \mathcal{X}_v$ which accounts for the multiple views of the object. Now, we consider the setting in which there is a large amount of objects and that most of them is likely to be remain unlabeled.

Multiview learning relies on two main assumptions: (1) The predictors in each view must agree on the majority of labels. (2) Each view is considered to be independent from the others, conditionally on labeled data. This second assumption is actually the one motivating most of the work on the multiview setting: we want to provide theoretical justification to the *heuristic* idea that (2) allows for high-performance results since two compatible classifiers trained in independent views are unlikely to agree on a mislabeled item. Following up on the early work of Blum and Mitchell (1998) on learning from both labeled and unlabeled data, several authors have developed a number of exciting achievements on this topic (see Blum and Chawla (2001), Sridharan and Kakade (2008), Weston et al. (2005), Zhou et al. (2003), Chapelle et al. (2006)). In florina Balcan and Blum (2005), the authors propose a nice theoretical PAC-model for semi-supervised learning where multiview learning appears as a special case. Then in Rosenberg and Bartlett (2007), these results are applied to a simple

two-view learning problem and explicit bounds on the Rademacher complexity of the class of predictors are computed. The issue which was tackled was to explain how consistency between views affects the performance. More recently, another general framework has been presented in Rosenberg (2008), very similar to this work. The RKHS (and kernels) of each view are tied together to compute the general kernel associated to the corresponding objective function, which then enables one to apply standard RKHS techniques. For instance, this gives an indirect alternative proof for the Rademacher complexity bound of Theorem 8.1. We consider a slightly less general objective function, with separate smoothness and agreement terms, that potentially enables more "interpretation" of the results, which is important when designing new parameter selection methods. Note that the full multiple view setting has also been considered in the work of Brefeld et al. (2006) for the specific co-RLS objective function, where the authors focused on the time complexity to find a closed form solution, and no Rademacher complexity analysis was provided. Finally, some local results were mentioned in Rosenberg (2008), Sindhwani and Rosenberg (2008), but the L_2-diameter analysis (see Theorem 8.3) has not been considered so far in any concurrent work for this setting, and goes beyond anything considered in Rosenberg and Bartlett (2007).

This chapter is organized as follows: in Section 2 we introduce our framework and define the objective function. The Section 3 is devoted to our result on the Rademacher complexity control. The main theorem is stated in section 3.2 together with its proof. Section 3.3 deals with the asymptotic behavior of the bounds. Then Section 4 shows some experiments on three toy examples. Finally, Section 5 details a stability-based selection procedure as well as a bound on the L_2 local diameter of our class of functions.

2 Setup for multiview semi-supervised learning

Our approach follows the popular method of penalized empirical risk minimization in RKHS, which leads to computable data-dependent terms of the objective function. Now, designing such a function is the frontier between science and art. The nature of the problem described in the introduction leads to different sources of penalization which we describe in this section. Namely, coming from semi-supervised learning, where part of the data is unlabeled, a smoothness penalization takes care of the structure (manifold) depicted by the data, as in Belkin et al. (2005). For the multiview setting (see Blum and Mitchell (1998)) we use another penalization - an agreement term - since two representations of the same object have to be given the same label.

We assume that each view is a RKHS, and write $\mathcal{F}^{(v)}$ for the v-th view of functions on $\mathcal{X}^v$ with value in $\mathcal{Y}$, the label set. Note that the label set is the same for all views. For simplicity, we will assume that $\mathcal{Y} = [-1, 1]$. We consider both smoothness and agreement terms in the same objective function since we deal with both ideas. This is an important improvement on the work of Sindhwani et al. (2005) where objective functions with only one of each term (and corresponding algorithms) are considered. The goal we pursue here is of

unifying algorithms instead of comparing them. In this section, we provide the notations and definitions of the penalty terms involved.

2.1 Semi-supervised regularization

The aim of semi-supervised learning is to work with data a part of which is labeled and the other part unlabeled. This setup is situated between classification and clustering theory. Specifically, the labeling provides a clue to design an objective function which is generally the main problem of clustering (where there is no labeling, and thus no objective truth), and the unlabeled part drives us throw the path of structure detection in the data, for instance by considering the data points depict a specific manifold. In the sequel, we will consider a batch of n i.i.d. data points, sampled according to a distribution $\mathbb{P}$, l of which are labeled, and the remaining $u = n - l$ are unlabeled. We index with $i \in \{1, \dots, l\}$ the labeled points $(x_i)_{i \leqslant l}$ together with their label $(y_i)_{i \leqslant l}$, and with $i \in \{l+1, \dots, l+u\}$ the unlabeled points $(x_i)_{l < i \leqslant l+u}$. We shall use exponents v for the representation in each view and indices i for the corresponding data point. Thus, each x_i is the representation in all views $x_i = (x_i^{(1)}, ..., x_i^{(V)})$.

We express the search for structure with a Smoothness term. A natural choice is the one using the graph-Laplacian, as explained for instance in Kondor and Lafferty (2002), Smola and Kondor (2003), Belkin et al. (2004), Ando and Zhang (2007), and where different operators based on this notion are used. The idea underneath the use of the graph-Laplacian is to consider that the data points depict a manifold (see Belkin et al. (2005)), and thus the graph-Laplacian can be seen as a discrete version of the Laplace-Beltrami differential operator. Assuming we have for each view v a similarity graph given by its adjacency matrix $W^{(v)}$, then the (unnormalized) graph-Laplacian is defined as $L^{(v)} = D^{(v)} - W^{(v)}$, where $D^{(v)}$ is the diagonal matrix $D^{(v)}_{i,i} = \sum_j W^{(v)}_{i,j}$. Other interesting choices are the symmetrical or random walk normalized graph-Laplacian. We intuitively want that each function $f^{(v)} \in \mathcal{F}^v$ be smooth w.r.t similarity structures in all views. Thus, here we will use a global smoothness operator based on the weighted average graph-Laplacian $L = \sum_{v=1}^{V} \alpha_v L^{(v)}$ with weights α summing to 1.

Definition 8.1 *For $f = (f^{(1)}, ..., f^{(V)}) \in \prod_{v=1}^{V} \mathcal{F}^{(v)}$, we define:*

$$Smoothness(f) = \sum_{v=1}^{V} \gamma_v \, \boldsymbol{f}^{(v)T} \, L \, \boldsymbol{f}^{(v)}, \text{ where}$$

- $\gamma = (\gamma_1, ..\gamma_V) \geqslant 0$ *meaning that each component is positive.*
- *L is defined based on $L^{(v)}$, the graph-Laplacian corresponding to the vth view: $L = \sum_{v=1}^{V} \alpha_v L^{(v)}$ with $\sum_{v=1}^{V} \alpha_v = 1$.*
- *$\boldsymbol{f}^{(v)}$ is the vector $(f^{(v)}(x_1^{(v)}), ..., f^{(v)}(x_{l+u}^{(v)}))^T$.*

2.2 Learning with RKHS

The main term of the objective function is the loss, which quantifies the classification errors on labeled points. This term has been widely studied along the past decade, and led to interesting theory involving convex risk minimization and φ-risks (see Bartlett et al. (2003) for a survey).

Classical losses for binary classification are for instance the square loss $\mathcal{L}(u,y) = (u-y)^2$, with y a label, the hinge loss $\mathcal{L}(u,y_i) = \max(0, 1-uy_i)$ or the logit loss $\mathcal{L}(u,y_i) = \log(1+e^{-uy_i})$. Based on this loss for one-view classifier, we define the loss of a *multiview classifier* f with the one of each $f^{(v)}$ in each view. More precisely:

Definition 8.2 *For $f = (f^{(1)}, \ldots, f^{(V)}) \in \prod_{v=1}^{V} \mathcal{F}^{(v)}$ and a sample $(x_i, y_i)_{1 \leqslant i \leqslant l}$, we define the loss of the multiview classifier f to be*

$$Loss(f) = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(f, x_i, y_i),$$

where we allow for various definitions for $\mathcal{L}(f,x,y)$, like for instance

$$\mathcal{L}(f,x,y) \stackrel{\text{def}}{=} \frac{1}{V} \sum_{v=1}^{V} \mathcal{L}(f^{(v)}(x^{(v)}), y) \;\; or \;\; \mathcal{L}(f,x,y) \stackrel{\text{def}}{=} \mathcal{L}(\frac{1}{V} \sum_{v=1}^{V} f^{(v)}(x^{(v)}), y)\,.$$

The choice of the loss will have an impact on the design of algorithm but our analysis covers a wide variety of possible losses. In the experiments, we use the square loss for one-view classifier, and the first definition for $\mathcal{L}(f,x,y)$.

We now introduce real-valued *decision functions*, that are defined to be functions φ of the form $x \mapsto \frac{1}{V} \sum_{v=1}^{V} f^{(v)}(x^{(v)})$ where $f^{(v)} : \mathcal{X}^{(v)} \to \mathcal{Y}$ is a classifier. We assume that each predictor $f^{(v)}$ lives in an RKHS $\mathcal{F}^{(v)}$ with kernel $K^{(v)}$ and associated representation function $k_v(.,.)$. Thanks to the representer theorem (see Schölkopf et al. (2001)), we will be able to restrict only to functions $f^{(v)} \in \mathcal{L}_v = \text{span}\{k_v(x_i^{(v)}, .)\}_{i=1}^{l+u} \subset \mathcal{F}^{(v)}$. We also denote by $\mathcal{F}$ the product space of the views $\mathcal{F}^{(v)}$ and by $\mathcal{L} \subset \mathcal{F}$ the product space of the spans. With each view comes a natural norm (the norm of the rkhs) which is interpreted as a measure of the complexity of the functions. We denote it by $||.||_v$ for the v-th view. Complexity penalization in RKHS has been studied for instance in Steinwart et al. (2006) or Blanchard et al. (2008).

The complexity term for the multi-function $f \in \mathcal{F}$ is thus stated according to the following definition:

Definition 8.3 *For $f = (f^{(1)}, ..., f^{(V)})$, we define:*

$$Complexity(f) = \sum_{v=1}^{V} \lambda_v ||f^{(v)}||_v^2$$

where $\lambda \in \mathbb{R}^V$ is a positive vector.

2.3 Multiple view co-regularization

The multiview setting comes with the need for compatibility between different views, since all representations of the same object must share the same label. Thus, it is natural to think that since there is a restriction on the search space, multiview learning will provide good generalization results, and indeed this is the case in numerical experiments (e.g. Belkin et al. (2005)). The work of Rosenberg and Bartlett (2007) and Sindhwani et al. (2005) aims at giving theoretical explanation and better understanding for this phenomenon. Our work follows the same red-line, generalizing and continuing this work.

The need for compatibility between the $f^{(v)}$ is conveyed by a so called Agreement term. First, for a number of $V = 2$ views, we can think of a term like $\sum_{i=1}^{n}[f^{(1)}(x_i^{(1)}) - f^{(2)}(x_i^{(2)})]^2$, with a l^2 penalty when giving two different labels to the same object. Note that here this penalty is mainly guided by convenience and easy matrix-wise formulation. We can also split this term between labeled and unlabeled points with some weights. Now, extending the notions presented in Sindhwani et al. (2005), and generalizing this to V views, we propose the following definition:

Definition 8.4 *For $f = (f^{(1)}, ..., f^{(V)}) \in \prod_{v=1}^{V} \mathcal{F}^{(v)}$, we define:*

$$\textit{Agreement}(f) = \mathcal{C}^L(f) + \mathcal{C}^U(f)\,,$$

$$\textit{with } \mathcal{C}^L(f) = \sum_{v_1 \neq v_2} c^L_{v_1,v_2} \sum_{i=1}^{l} [f^{(v_1)}(x_i^{(v_1)}) - f^{(v_2)}(x_i^{(v_2)})]^2\,,$$

$$\textit{and } \mathcal{C}^U(f) = \sum_{v_1 \neq v_2} c^U_{v_1,v_2} \sum_{i=l+1}^{l+u} [f^{(v_1)}(x_i^{(v_1)}) - f^{(v_2)}(x_i^{(v_2)})]^2\,.$$

For normalization purpose, we assume that the positive coefficients $c^L_{v_1,v_2}, c^U_{v_1,v_2}$ are 0 whenever their value do not modify $\mathcal{C}^L(f)$ or $\mathcal{C}^U(f)$ (for instance when $f^{(v_1)}(x^{(v_1)}) = f^{(v_2)}(x^{(v_2)})$ $\mathbb{P}$-a.s.), and that $c^L = (c^L_{v_1,v_2})_{v_1,v_2}$, $c^U = (c^U_{v_1,v_2})_{v_1,v_2}$ are symmetric semi-definite positive. So as to avoid cumbersome terms in the proof, we also introduce when $v, w \in \{1, .., V\}$ the block-diagonal matrix $C_{v,w}$ with diagonal blocks $(c^L_{v,w})_{i=1..l}$ and $(c^U_{v,w})_{i=l+1..l+u}$, and then $C \in \mathbb{R}^{nV(V-1)\times nV(V-1)}$ the block-diagonal matrix with blocks $(C_{v,w})_{v,w}$.

2.4 Compound complexity penalties

The objective function and its associated minimization problem we consider in this work is finally written as:

- Compute:

$$f^* = \arg\min_{f\in\mathcal{F}} \{ \text{Loss}(f) + \text{Complexity}(f) + \text{Smoothness}(f) + \text{Agreement}(f) \} \quad (8.1)$$

- Output: $\varphi = \frac{1}{V}\sum_{v=1}^{V} f^{*(v)}$

This will be our minimization algorithm for multiple-view semi-supervised learning in this work. We note throughout this chapter $\theta = (\alpha, \lambda, \gamma, \mathcal{C})$ to refer to all the parameters appearing in the objective function.

The existence of a representer theorem for this setting comes from the fact that for any fix $f^{(2)}, ..., f^{(V)} \in \Pi_{v=2}^{V} \mathcal{F}^{(v)}$, $f^{*(1)}$ minimizes a loss function of the form

$$c_{f^{(2)},...,f^{(V)}}(f(x_1^{(1)}), y_1, ..., f(x_n^{(1)}), y_n) + g_{f^{(2)},...,f^{(V)}}(||f||_1)$$

w.r.t. f. Thus the representer theorem (see Schölkopf et al. (2001)) says that $f^{(1)} \in \mathcal{L}_1$, and we can iteratively apply the same argument to each component of f^*, leading eventually to $f^* \in \mathcal{L}$. See also Sindhwani and Rosenberg (2008) for the construction of one RKHS combining all the views.

Note that for specific choices of the parameters, we recover the former problems studied in previous papers:

- when γ and C are 0, i.e. when the Smoothness and Agreement terms disappear, then we recover Regularized Least Squares (RLS) in RKHS,
- when only $\gamma = 0$, then we have a Co-Regularized Least Squares (co-RLS) problem (see Sindhwani et al. (2005)),
- when Agreement is nonzero but diagonal (meaning c^L and c^U are diagonal), we obtain the formulation of the co-laplacian method (such as co-laplacian RLS and co-laplacian SVM, see Sindhwani et al. (2005)) ; indeed here, the predictors $f^{(v)}$ are decoupled, and thus the problem 8.1 amounts to solving for each v:

$$f^{(v)*} = \arg\min_{f^{(v)}\in\mathcal{F}^{(v)}} \text{Loss}(f^{(v)}) + \lambda_v ||f^{(v)}||_v^2 + \gamma_v \mathbf{f}^{(v)T} L \mathbf{f}^{(v)} \ .$$

Our first purpose in this chapter is to generalize both the work of Sindhwani et al. (2005) and Rosenberg and Bartlett (2007), by taking into account all penalty terms simultaneously and considering arbitrary many views.

3 Empirical complexity bound

This section is devoted to the control of the Rademacher complexity for our problem. We define our class of functions in subsection 3.1, then we state our main theorem in subsection 3.2, which is a bound on the Rademacher complexity of this class, together with its proof. The last subsection studies asymptotic behavior of the bounds.

We first state the following assumption from Rosenberg and Bartlett (2007), which is satisfied for instance by the square loss. This assumption enables to reduce the search space and to perform computations.

Assumption A1: We suppose that the loss functional satisfies $\text{Loss}(0, ..., 0) \leqslant 1$ where $(0, ..., 0)$ is the multi-predictor with constant output 0.

This is true for instance, for the functional defined with the square loss. Indeed $\forall i \quad y_i \in [-1, 1]$, we have $\text{Loss}(0, .., 0) = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{V}\sum_{v=1}^{V} y_i^2 \leqslant 1$.

3.1 Preliminaries

We now recall the definition of the Rademacher complexity for a class of functions, that is a useful empirical quantity in order to derive an excess risk bounds for the empirical minimizer of a loss function over a class of functions (Boucheron et al. (2005)), like the decision function φ.

Definition 8.5 (*Rademacher complexity*) *The Rademacher complexity of a class $\mathcal{G}$ for a sample $(x_1, ...x_n)$ is*

$$R_n(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i g(x_i)\right|\right]$$

where $(\sigma_i)_{i\leqslant n}$ are Rademacher i.i.d. random variables defined by

$$\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}\,.$$

In our case, the final predictor φ is just a combination of the predictors $f^{(v)}$ on each view. Under assumption A1, φ belongs to the class $\mathcal{J} = \mathcal{J}(\theta, 1)$ defined by:

$$\mathcal{J}(\theta, r) = \left\{x \to \frac{1}{V}\sum_{v=1}^{V} f^{(v)}(x^{(v)}) : \quad (f^{(1)}, .., f^{(V)}) \in \mathcal{H}(\theta, r)\right\}$$

where $\mathcal{H}(\theta, r)$ is the class of multi-predictors f, with total penalty bounded by r:

$$\mathcal{H}(\theta, r) = \{f \in \mathcal{L} : \text{Complexity}(f) + \text{Smoothness}(f) + \text{ Agreement}(f) \leqslant r\}$$

3.2 Explicit Rademacher complexity bound

Before stating the main theorem of this section, we need to introduce some matrix notations.

Block-wise notations First Id_n is the identity of $\mathbb{R}^n$ and $0_{u,l}$ is the zero matrix of size $u \times l$. Since we have matrices and vectors corresponding to each view, we use a block-wise notation. Roughly speaking, each block will denote one view. Thus, for any given n_1, n_2 and a matrix $A^{(v)} \in \mathbb{R}^{n_1 \times n_2}$, we write $\overline{A}$ for the block-diagonal matrix with blocks $A^{(v)}$, $v = 1..V$ (of size $n_1 V \times n_2 V$), and similarly $\underline{A}$ the block-row matrix of size $n_1 V \times n_2$. Now suppose we want to multiply block-wise each block $A^{(v)}$ by the v-th component of a vector $\lambda \in \mathbb{R}^v$. To do that we introduce $\tilde{\lambda}$ the block-diagonal matrix of size $n_1 V \times n_1 V$ with blocks $\lambda_v \mathrm{Id}_{n_1}$. Since it is generally clear that we want to multiply the v-th block with the v-th component, we forget the n_1 in the notations. This allows us to focus on one view.

Labeled and unlabeled data We decompose each kernel matrix K according to labeled and unlabeled data, with first rows K_L for the labeled part and the last ones K_U for the unlabeled part. We similarly introduce the projection matrix of labeled data Π; we have

$$K = \begin{pmatrix} K_L \\ K_U \end{pmatrix} \in \mathbb{R}^{n \times n} \text{ and } \Pi = \underline{\begin{pmatrix} \mathrm{Id}_l \\ 0_{u,l} \end{pmatrix}} \in \mathbb{R}^{nV \times l}.$$

Agreement Finally, to handle our agreement term, we need to compare views pairwise. To do that, we thus introduce a special matrix $\delta \in \mathbb{R}^{nV(V-1) \times nV}$, block-line defined with blocks

$$(0 \ \ldots \ 0 \ \mathrm{Id}_n \ 0 \ \ldots \ 0 \ -\mathrm{Id}_n \ 0 \ \ldots \ 0)$$

with identity matrices at position v_1 and $v_2 \neq v_1$.

Smoothness Finally, the smoothness term directly appears in the Rademacher bound through the matrix $B = (I + \tilde{\lambda}^{-1} \tilde{\gamma} L_I \overline{K})^{-1}$, where L_I is the diagonal block matrix with all V blocks equaled to L. Note that we would have $\tilde{\alpha}\overline{L}$ instead of L_I if we have used the graph Laplacian of each view in the smoothness term instead of the average Laplacian L.

Thanks to the previous notations, we can now state our main theorem, which shows an explicit upper and lower data-dependent bound for the Rademacher complexity of our class of functions.

Theorem 8.1 (*Rademacher complexity bound*) *The Rademacher complexity of the class of decision functions, under assumption A1, is bounded above and below by*

$$\frac{1}{2^{1/4}} \frac{2b}{Vl} \leqslant R_l(\mathcal{J}) \leqslant \frac{2b}{Vl}$$

where

$$b^2 = tr(B\tilde{\lambda}^{-1}\Pi\underline{K_L^T}^T) - tr(J'^T(I+M')^{-1}J')$$

with

- $B = (I + \tilde{\lambda}^{-1}\tilde{\gamma}L_I\overline{K})^{-1} \in \mathbb{R}^{nV\times nV}$
- $J' = \sqrt{C}\delta\tilde{\lambda}^{-1}B^T\underline{K_L^T} \in \mathbb{R}^{nV(V-1)\times l}$
- $M' = \sqrt{C}\delta\overline{K}B\tilde{\lambda}^{-1}\delta^T\sqrt{C} \in \mathbb{R}^{nV(V-1)\times nV(V-1)}$

Before stating the proof, let us make some comments. Note that b is explicit, and that it consists of two terms. The first term only depends on unlabeled data when Smoothness is null, and contains no co-regularization term. The second term corresponds to the idea that there is a reduction in complexity of the space. Indeed, in section 3.3, we give some results about the behavior of b enforcing this idea. Note that the special shape of this bound $\frac{1}{l}\sqrt{tr(\kappa)}$ that reminds usual Rademacher bounds for balls in RKHS is not pure coincidence : as pointed out in Sindhwani and Rosenberg (2008), this term is connected to a specific RKHS-norm induced by the parameters and data on the space. The corresponding kernel may even be built explicitly and gives an alternative indirect proof to the above theorem, seeing $\mathcal{J}$ as a ball in this RKHS.

Example: if $\gamma = 0$, then the b term reduces to

$$b^2 = \sum_{v=1}^{V}\lambda_v^{-1}tr(K_{L,L}^{(v)}) - tr(J^T(I+M)^{-1}J)$$

where J is the block-diagonal matrix with blocks $\sqrt{C_{v_1,v_2}}[\lambda_{v_1}^{-1}K_{L,LU}^{(v_1)} - \lambda_{v_2}^{-1}K_{L,LU}^{(v_2)}]$ and M the block-diagonal matrix with blocks $C_{v_1,v_2}[\lambda_{v_1}^{-1}K^{(v_1)} + \lambda_{v_2}^{-1}K^{(v_2)}]$ ($K_{L,L}$ and $K_{L,LU}$ are the sub-matrices of K of respective size $l\times l$ and $l\times(l+u)$ corresponding to the labeled and unlabeled points). This mean that in the special case of $V = 2$ views and when c^L is the 0 matrix, we recover exactly the previous known bound of Rosenberg and Bartlett (2007).

We now prove Theorem 8.1. The proof technique is similar to Rosenberg and Bartlett (2007) and we use it to extend their result in order to cover the compound regularization penalty in the case of an arbitrary number of views.

Proof: The proof is in five steps. First, we show that under assumption A1, the solution lives in the space $\mathcal{H}$. Then, we use the representer theorem to have a matrix reformulation of what this means. Then, in order to compute the Rademacher complexity, we need a matrix to be invertible, which we obtain by a reformulation using classical invariance properties of the kernel function. Eventually, we rewrite the Rademacher complexity in terms of the initial data.

Step 1. Solution Space. First, we reduce the search space to the space of spans over the kernel matrices involving only the data points, intersected with the space of predictor with a penalty term upper bounded by 1.

Lemma 8.1 *Under assumption A1, the solution of the minimization problem 8.1 belongs to the set* $\mathcal{L} \cap \mathcal{H}$.

Proof: We call Q the functional to be minimized. we can write Q as the sum of two terms: $Q(f) = \text{Loss}(f) + \Pi(f)$, and moreover for $0 \in \mathcal{F}$, the null multiview predictor, we have $Q(0) = \text{Loss}(0)$, thus under assumption A1, $\inf Q \leqslant 1$. But since all terms of Q are non negative, the solution is in the set $\mathcal{H}$. Finally, the fact that $f^* \in \mathcal{L}$ is just an application of the representer theorem which holds for our setting. □

Step 2. Matrix formulation. If $f \in \mathcal{L} \cap \mathcal{H}$, then thanks to the representer theorem, we can write its component in each view $f^{(v)} = f^{(v)}_{\alpha^{(v)}} = \sum_{i=1}^{n} \alpha_i^{(v)} k_v(., x_i^{(v)})$, with parameter $\alpha^{(v)} \in \mathbb{R}^n$. Thus, after easy computation, a matrix reformulation of $f \in \mathcal{L} \cap \mathcal{H}$ is:

$$f \in \{(f_{\alpha^{(1)}}, ..., f_{\alpha^{(V)}}) : \quad \underline{\alpha}^T N \underline{\alpha} \leqslant 1\}$$

where $\alpha \in \mathbb{R}^{nV \times 1}$, the data-dependent N square matrix is

$$N = \tilde{\lambda}\overline{K} + \tilde{\gamma}\text{Diag}\left(K^{(1)}LK^{(1)} \dots K^{(V)}LK^{(V)}\right) + \sum_{v_1 \neq v_2} K_C^{v_1,v_2}$$

The notation $\text{Diag}(v_1....v_k)$ is a shortcut for the square matrix with diagonal blocks $v_1, ..., v_k$ on the diagonal.

Finally, the last term of N is the following agreement matrix:

$$K_C^{v_1,v_2} = \begin{pmatrix} 0 \\ \vdots \\ K^{(v_1)} \\ \vdots \\ -K^{(v_2)} \\ \vdots \\ 0 \end{pmatrix} C_{v_1,v_2} \begin{pmatrix} 0 \\ \vdots \\ K^{(v_1)} \\ \vdots \\ -K^{(v_2)} \\ \vdots \\ 0 \end{pmatrix}^T .$$

With this matrix formulation, our bounding problem can be seen as an optimization problem under quadratic constraint. Indeed, since $\mathcal{H}$ is symmetrical, one can write R_l as:

$$R_l(\mathcal{J}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{L} \cap \mathcal{H}} \frac{2}{lV} \sum_{i=1}^{l} \sigma_i \sum_{v=1}^{V} f^{(v)}(x_i^{(v)}) \right]$$

and this is again $R_l(\mathcal{J}) = \frac{2}{lV}\mathbb{E}_\sigma \sup_{\alpha;\alpha^T N\alpha \leqslant 1} \alpha^T \underline{K_L^T}\sigma$, with $\sigma = (\sigma_1, \dots, \sigma_l)^T \in \mathbb{R}^{l \times 1}$.

Step 3. Basis change. In order to compute the supremum, we use the following lemma:

Lemma 8.2 *If M is a symmetric positive definite (spd) matrix, then*

$$\sup_{\alpha:\alpha^T M\alpha \leqslant 1} v^T\alpha = ||M^{-1/2}v||$$

Proof: From Karush-Kuhn-Tucker conditions, with dual variable $\lambda \geqslant 0$, we have : $v = 2\lambda M\alpha$ and $(\alpha^T M\alpha - 1)\lambda = 0$. Since M is invertible, we deduce that: $2\alpha = \lambda^{-1}M^{-1}v$ and $v^T M^{-1}v = 4\lambda^2$. Now since M is spd, we have $v^T M^{-1}v = ||M^{-1/2}v||^2$. Thus the maximal value of $v^T\alpha$ is reached for $\frac{1}{2\lambda}||M^{-1/2}v||^2 = ||M^{-1/2}v||$. □

However, our matrix N may not have full rank. Thus, we use the eigen decomposition of $K^{(v)}$: $P^{(v)T}K^{(v)}P^{(v)} = \Sigma^{(v)}$ where $\Sigma^{(v)}$ is the diagonal matrix of the $m^{(v)}$ non-zero eigenvalues of $K^{(v)}$, $P^{(v)}$ is rectangular with size $n \times m^{(v)}$ and $\Sigma^{(v)}$ is $m^{(v)} \times m^{(v)}$. Now, $\Sigma^{(v)}$ is invertible. We write $M = \sum_{v=1}^{V} m^{(v)}$.

In order to use this, we introduce $\alpha_{//}^{(v)}$, the projection of the $\alpha^{(v)}$ on the subspace associated to the rows of the $K^{(v)}$. Note that the quadratic form $\underline{\alpha}^T N\underline{\alpha}$ is left unchanged under this projection, which allows us write $\alpha_{//}^{(v)} = P^{(v)}a^{(v)}$, introducing the vector $a^{(v)}$. We then rewrite our set in terms of the new parameters:

$$\mathcal{H} = \{(f_{a^{(1)}}, ..., f_{a^{(V)}}) : \quad \underline{a}^T T\underline{a} \leqslant 1\}$$

where $T = \tilde{\lambda}\overline{\Sigma} + \tilde{\gamma}\overline{\Sigma}\tilde{L}\overline{\Sigma} + \sum_{v_1 \neq v_2=1}^{V} R_C^{v_1,v_2} R_C^{v_1,v_2\,T}$, $\tilde{L} = \overline{P^T}L_I\overline{P} \in \mathbb{R}^{M\times M}$ and

$$R_C^{v_1,v_2} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & P^{(v_1)T} & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & P^{(v_2)T} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ K^{(v_1)} \\ 0 \\ \vdots \\ -K^{(v_2)} \\ \vdots \\ 0 \end{pmatrix} \sqrt{C_{v_1,v_2}}$$

Step 4. Rademacher complexity bounds. Now, T is invertible. If we split each $P^{(v)}$ according to the label and unlabeled part to get $(P_L^{(v)T}P_U^{(v)T})^T$ and introduce $W = \underline{K_L^T}^T\overline{P} = \underline{P_L^T}^T\overline{\Sigma}$, then using the lemma 8.2 we can write:

$$R_l(\mathcal{J}) = \frac{2}{lV}\mathbb{E}_\sigma||T^{-1/2}W^T\sigma||$$

Following the proof of Rosenberg and Bartlett (2007), we now apply the Kahane-Khintchine inequality to get $\frac{1}{2^{1/4}}\frac{2b}{lV} \leqslant R_l(\mathcal{J}) \leqslant \frac{2b}{lV}$, where $b^2 = \mathbb{E}_\sigma||T^{-1/2}W^T\sigma||^2$. But using the definition of the norm, this term can be also written $||WT^{-1}W^T\sigma\sigma^T|| = tr[WT^{-1}W^T]$.

Thus, we need to compute the term T^{-1}. To do that, we use the decomposition $T = (A + UU^T)$ and the following Sherman-Morrison-Woodbury formula, seeGolub and Van Loan (1996):

Lemma 8.3 *Provided that the inverses exist, we have:*

$$(A + UU^T)^{-1} = A^{-1} - A^{-1}U(I + U^T A^{-1}U)^{-1}U^T A^{-1} \ .$$

Identifying A and U, we find $A = \tilde{\lambda}\overline{\Sigma} + \tilde{\gamma}\overline{\Sigma}\tilde{L}\overline{\Sigma} = \tilde{\lambda}\overline{\Sigma}(I + \tilde{\lambda}^{-1}\tilde{\gamma}\tilde{L}\overline{\Sigma})$. A is invertible since $\tilde{L}$ is spd, and Σ is invertible and diagonal. On the other hand $U = J_U\sqrt{C}$ where J_U is of size $M \times nV(V-1)$, defined block-wise $[J_U^{1,2}|...|J_U^{V,V-1}]$ where $J_U^{v,w}$ is the product of $\text{Diag}\left(0\dots\Sigma^{(v_1)}0\dots\Sigma^{(v_2)}\dots 0\right)$ with $\left(0\dots P^{v_1}0\cdots - P^{v_2}\dots\right)^T$

Step 5. Rademacher bound in terms of the initial data. Now, we can express b in terms of the initial data by replacing the terms A and U in T^{-1} with their corresponding value. Putting this in $tr(WT^{-1}W^T)$, and after some careful but easy computations, we find the bounds of the Theorem. Indeed, by the previous lemma, $b^2 = tr(WA^{-1}W^T) - tr(WA^{-1}U(I+U^TA^{-1}U)^{-1}U^TA^{-1}W^T)$. The first term is again, using the definition of A and W : $tr(\underline{K_L}^T\overline{P}(I + \tilde{\lambda}^{-1}\tilde{\gamma}\tilde{L}\overline{\Sigma})^{-1}\tilde{\lambda}^{-1}\overline{\Sigma}^{-1}\overline{\Sigma}\underline{P_L}^T)$, which is also $tr(\overline{P}(I + \tilde{\lambda}^{-1}\tilde{\gamma}\tilde{L}\overline{\Sigma})^{-1}\tilde{\lambda}^{-1}\underline{P_L^TK_L^T}^T)$. Thus we recognize here the first term of b^2 in the Theorem. For the second term, since

$$U^TA^{-1}U = \sqrt{C}^TJ_U^T(I + \tilde{\lambda}^{-1}\tilde{\gamma}\tilde{L}\overline{\Sigma})^{-1}\tilde{\lambda}^{-1}\overline{\Sigma}^{-1}J_U\sqrt{C}$$
$$U^TA^{-1}W^T = \sqrt{C}^TJ_U^T(I + \tilde{\lambda}^{-1}\tilde{\gamma}\tilde{L}\overline{\Sigma})^{-1}\tilde{\lambda}^{-1}\overline{\Sigma}^{-1}\overline{P}^T\underline{K_L}^T,$$

we identify these two terms respectively to M' and J', which concludes the proof. □

3.3 Asymptotics

The expression for b contains the different penalty parameters of the problem, which are contained in the vector $\theta = (\alpha, \lambda, \gamma, \mathcal{C})$. We recall that the parameter α appears in the graph-Laplacian, λ in the Complexity term, γ in the Smoothness term and finally $\mathcal{C}$ in the Agreement term. Note that the number of parameters grows with $O(V^2)$. It is interesting to understand how the bound on the Rademacher complexity behaves when the parameters vary.

More agreement reduces space complexity. We are mainly interested in the matrix C which controls the co-regularization. The second term appearing in the expression of b^2 depends on the co-regularization (matrix) parameter $\mathcal{C}$. To see how constraint is the space when using bigger penalization, we introduce the following $\Delta(C) = tr(J'^T(I + M')^{-1}J')$. In order to highlight the dependency in C, we can rewrite it (provided C^{-1} exists) as:

$$\Delta(C) = tr(J_1^T(C^{-1} + M_1)^{-1}J_1)$$

where J_1 and M_1 are defined likewise J' and M' but without the matrix C, i.e. $J_1 = \delta\tilde{\lambda}^{-1}B^T\underline{K_L}^T$ and $M_1 = \delta\overline{K}B\tilde{\lambda}^{-1}\delta^T$.

One interesting case is thus when the eigenvalues of C increase to $+\infty$, since the term $\Delta(C)$ indeed tends to the limit quantity:

$$\Delta_\infty = tr(\underline{K_L}^TB\tilde{\lambda}^{-1}\delta^T(\delta\overline{K}B\tilde{\lambda}^{-1}\delta^T)^{-1}\delta\tilde{\lambda}^{-1}B^T\underline{K_L}^T),$$

which can be rewritten $\Delta_\infty = tr(B\tilde{\lambda}^{-1}\Pi_l\underline{K_L}^{T^T})$, thus showing that $b^2 \to 0$ in this case. Note that the fact that b decreases as the model gets more constraint is coherent with the intuition

of multiview learning. Note also that by convention for C, for two identical views v_1 and v_2, the corresponding agreement parameter c_{v_1,v_2} is 0 and doesn't participate in the complexity decrease.

Other parameters. From the previous formula, we deduce that similarly, b^2 tends to zero whenever $||\gamma||$, or $||\lambda|| \to \infty$. On the other hand, when the penalty terms go to 0, i.e. when the constraint on the space vanishes, we have a completely different behavior. Indeed, we see that if $C = 0$ then $\Delta(C) = 0$. The case where $\gamma = 0$ has been considered by Rosenberg and Bartlett (2007), and finally, when only $\lambda = 0$, b^2 has the following expression, provided every term appearing in its expression is finite and defined:

$$b^2 = tr(\Pi_l^T L_I^{-1} \tilde{\gamma}^{-1} \Pi_l) - tr(\Pi_l^T L_I^{-1} \tilde{\gamma}^{-1} \delta^T (C^{-1} + \delta L_I^{-1} \tilde{\gamma}^{-1} \delta^T)^{-1} \delta L_I^{-1} \tilde{\gamma}^{-1} \Pi_l)$$

Note that when both γ and λ tend to 0, the previous bound may tend to ∞ even in simple cases (ex: $C = 0$), which is coherent with the intuition and shows that we do not have the same behavior at all. Note also that the dependency with V is hidden here in the trace.

4 Experiments

We have performed some toy simulations to see the flexibility of this general algorithm and the results are promising. Based on only very few labeled points, we can always recover perfect labeling of the data, even on one challenging dataset on which all previous multi-view algorithms (Co-Laplacian and Co-RLS) performs badly. There is no magic in this and we briefly describe what happens. For completeness, we first give hints how to solve the minimization problem in two particular cases: differentiable loss, and hinge loss.

4.1 Implementation

Recall that since a representer theorem holds for our setting, the solutions of the problem 8.1 can be written $f^{(v)}(x^{(v)}) = \sum_{i=1}^{l+u} \alpha_i^{(v)} K^{(v)}(x^{(v)}, x_i^{(v)}) = K^{(v)}_{x^{(v)}} \alpha^{(v)}$.

Differentiable loss: Suppose that the loss function is differentiable (for example, this happens for the quadratic loss), then the following theorem gives a solution to 8.1:

Theorem 8.2 (*Solution to the multiview learning program*) *Assuming that the loss function satisfies $\nabla_{\alpha^{(v)}} Loss(f^{(v)}) = 2K^{(v)} A^{(v)} \alpha^{(v)}$, then the solution of the problem 8.1 is given by the resolution of the linear system, where the $\alpha^{(v)}$ are the unknown vectors.*
$\forall v \in 1 \dots V$:

$$Y = [A^{(v)} + \lambda_v I + \gamma_v L K^{(v)}] \alpha^{(v)}$$
$$+2 \sum_{w=1}^{V} C_{v,w} (K^{(v)} \alpha^{(v)} - K^{(w)} \alpha^{(w)})$$

where $Y_i = y_i$ *for* $1 \leqslant i \leqslant l$ *and* $Y_i = 0$ *for* $l+1 \leqslant i \leqslant l+u$.

The proof is a straightforward application of usual algebra and thus is omitted here. The interested reader may notice that this system nicely contains as a special case the linear system of Sindhwani et al. (2005). It also immediately turns out that $\sum_{v=1}^{V} \alpha_i^{(v)} \lambda_v = 0$ for all $l+1 \leqslant i \leqslant l+u$.

Note that we can rewrite this system as $S\alpha = \tilde{Y}$ where S is an appropriate matrix, $\alpha = (\alpha^{(1)T}, .., \alpha^{(V)T})^T$ and $\tilde{Y} = (Y^T, .., Y^T)^T$, which can be solved using classical algebra. One has to be aware that S a priori enjoys no good properties, it is not positive in general and may have a very large conditioning number, so naively inverting the system gives highly unstable results.

Non-differentiable hinge loss: When Loss is not differentiable, an important case is SVMs (hinge loss), which cannot be solved with the previous linear system. Instead, this is done classically through linear programming, by introducing the so-called slack variables (see Bishop (2006)). The reader may want to derive explicitly the kernel associated to the underlying RKHS and then use classical SVM solvers. We refer for instance to Rosenberg (2008) to deal with this technical question. A complete derivation is given inBelkin et al. (2005) when $\gamma = 0$.

4.2 Toy examples

We have done some experiments on three toy examples, with only two views and two classes for simplicity.

- The first one is a classical two moons-two lines data set. This first dataset is easy, since in the second view the two lines are linearly separable, and in the first view the two-moons are almost separated.
- The second one is a more complex two spirals-two clouds data set. The complexity is due to the imbrication of the spirals, and actually, such a dataset has been generated to force the use of graph-laplacian. Note that one human being can not separate the two class without the information of the second view.
- The last one is a challenging one cross-two moons data set, which appears to fool the tested algorithms based on only one of the Smoothness or Agreement term. This is due to the cross which would be ambiguous with only one example of each class (thus we give two label examples).

These are depicted in figure 8.1. Since the less labeled object, the more heuristic is the definition of the "true" classes, in first approximation we refer here to human beings to say what are the true classes. Of course the definition of what a true class is is a real problem still

unsolved in the clustering community and we do not pretend here to solve it. In the first two data sets, a human only needs one label object of each class to recover the classes. For the last one, because of the cross which yields ambiguity, a human operator needs two objects in each class in order to remove ambiguity between the classes. Thus, we put exactly this number of labels point and run different algorithms. Thus the number l of labeled objects is very small compared to the total amount of data.

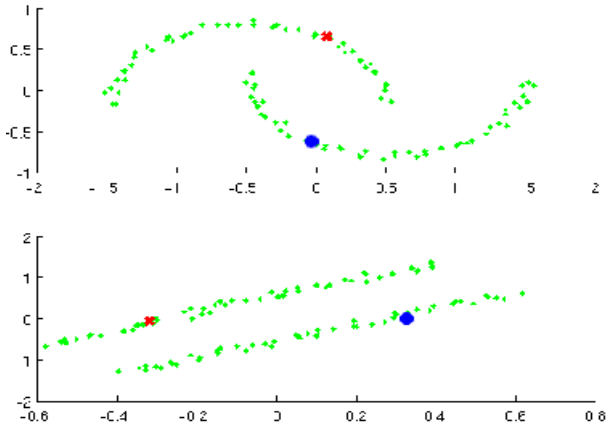
We use four different algorithms, all derived from our general algorithm but with different missing regularization terms. For each algorithm we use the same loss: the quadratic loss, which is differentiable. The first one is the classical RLS, for which Smoothness and Agreement are set to 0. As expected, this algorithm performs badly on all toy examples, even with appropriate choice of kernels. The second one is a co-RLS, with only Smoothness set to 0. Then we used a laplacian-based algorithm (co-laplacian), which outperform co-RLS on the tricky two spirals-two clouds data set, and finally we used the algorithm with none of the terms set to 0.

Graphical results are provided in figures 8.2,8.4,8.5. For each one, from top-left to right and down, we have the results of RLS, co-RLS, co-laplacian and the general algorithm (i.e. with no parameter set to zero). Graphical representation of the data conveys more information on the behavior of the algorithm than error rates. We also present a table with error rates of the different algorithms on the above toy examples. Due to the large number of parameters ($O(V^2)$), parameter selection is a tricky problem we address in the next section. For the experiments, we just tuned the parameters by hand trying to find the best results for each algorithm. While hand-tuning is surprisingly easy for some algorithms on specific toy examples (co-RLS for two moons-two lines, co-laplacian for two spirals-two clouds, general for one cross-two moons), it is quite difficult for others (e.g. co-laplacian for one cross-two moons), maybe due to numerical implementation, where we mean by difficult that the results are experimentally unstable w.r.t. a slight change of the parameters. Comparison of different selection procedures, although interesting by itself, is not the purpose of this work. And indeed, carrying out specific experiments on various toy as well as real problems for a fair and complete comparison (especially with recent impressive techniques, as the "slope" heuristics) goes beyond the scope of this work and should be discussed in a specific one.

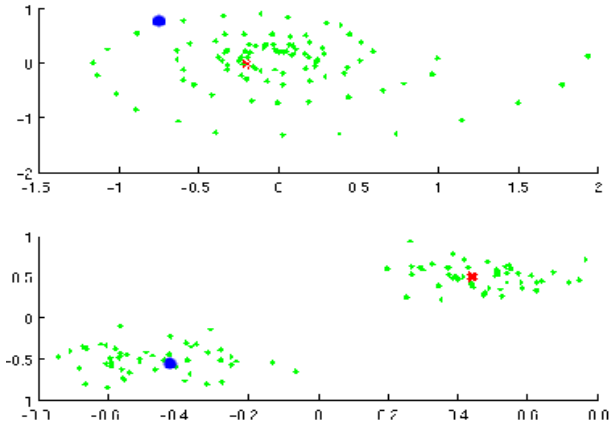
Finally, note that the choice of the kernels for each view is important, and we tried to use well-suited kernels (gaussian for clouds, linear for lines, ...).

5 Stability-based parameter selection

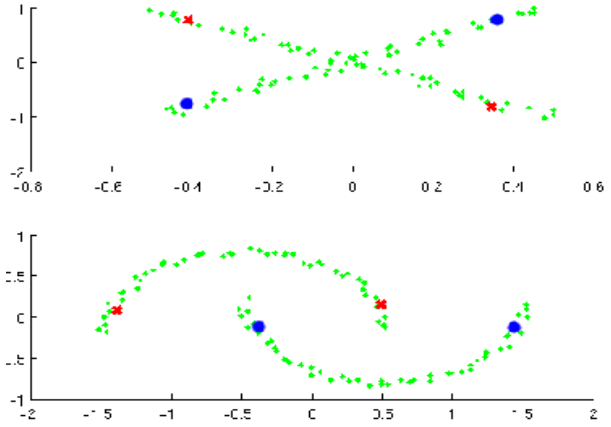
In the general algorithm we presented, the number of parameters is $O(V^2)$, so we need a selection procedure. In the previous results on toy examples, due to this large amount of parameters and since we artificially put some parameters to 0 to highlight some specific behavior, we used hand-tuned parameters. We now try to provide ways for automatic parameter selection.



Two moons-two lines



Two spirals-two clouds



One cross-two moons

Figure 8.1: Three toy data sets. Normal points is for unlabeled points, circle for class number one and cross for class number two. From left to right: Two moons (above)- two lines (below), with one labeled object in each class. Two spirals-two clouds, with again one labeled object in each class. One cross-two moons, with two labeled objects in each class

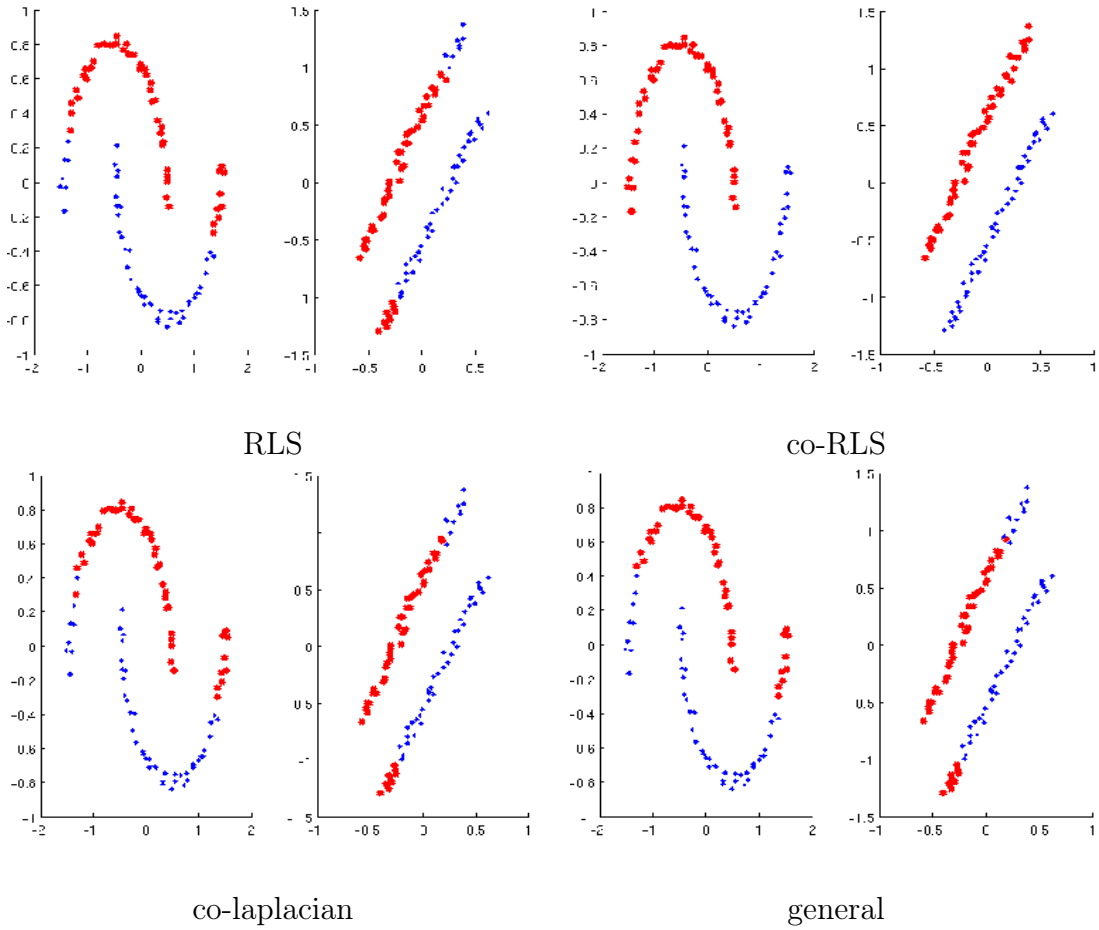


Figure 8.2: Two moons - two lines. For each algorithm (RLS, co-RLS, co-laplacian, general), the classes learned on the objects projected on each view. Here the co-RLS achieves exact classification

algo	dataset 1	dataset 2	dataset 3
RLS	0.455 ± 0.035	0.103 ± 0.024	0.379 ± 0.026
co-RLS	0.146 ± 0.071	0.103 ± 0.024	0.467 ± 0.025
co-Laplacian	0.242 ± 0.040	0.001 ± 0.004	0.510 ± 0.028
general	0.011 ± 0.015	0.322 ± 0.067	0.042 ± 0.071

Figure 8.3: Empirical misclassification errors for the above algorithms (one set of parameters per dataset, some possibly put to zero when specified to each algorithm), averaged over 1000 runs.

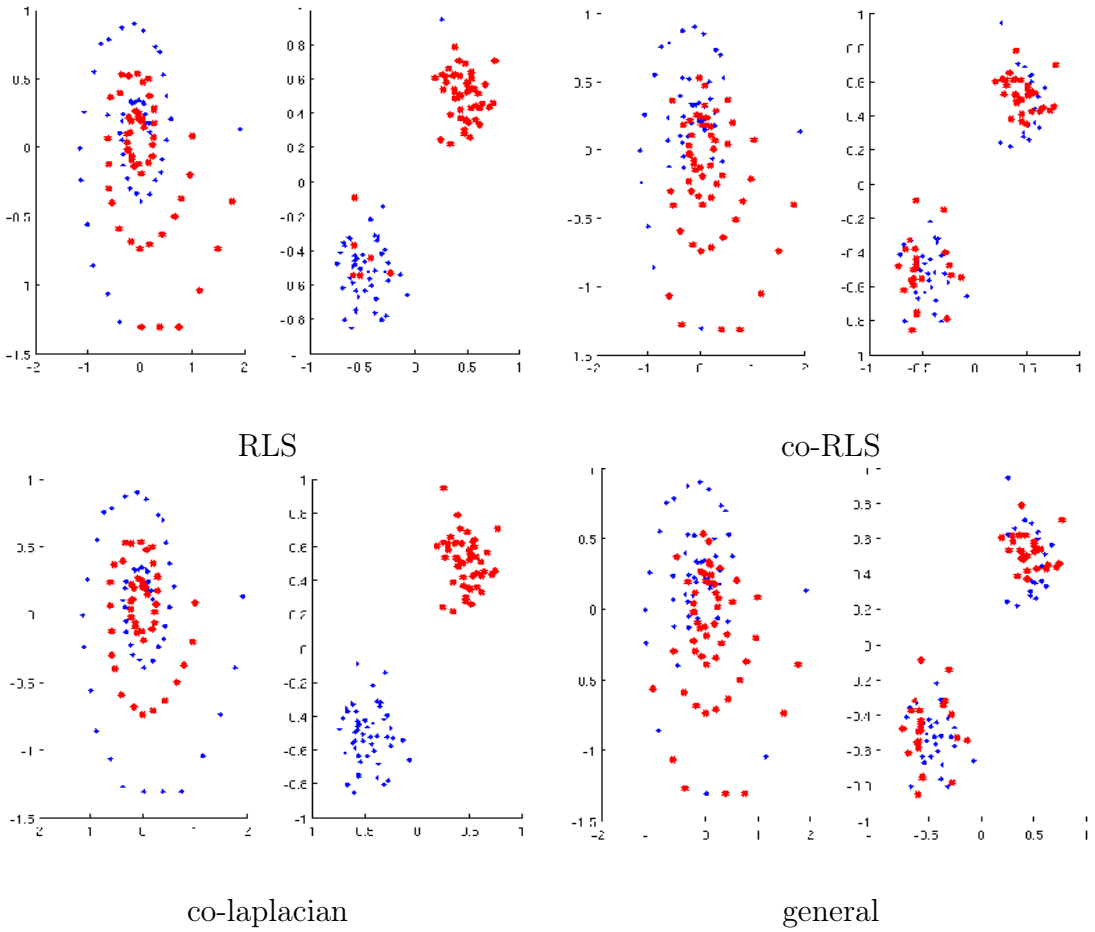


Figure 8.4: Two spirals - two clouds. For each algorithm (RLS, co-RLS, co-laplacian, general), the classes learned on the objects projected on each view. Here the co-laplacian achieves exact classification

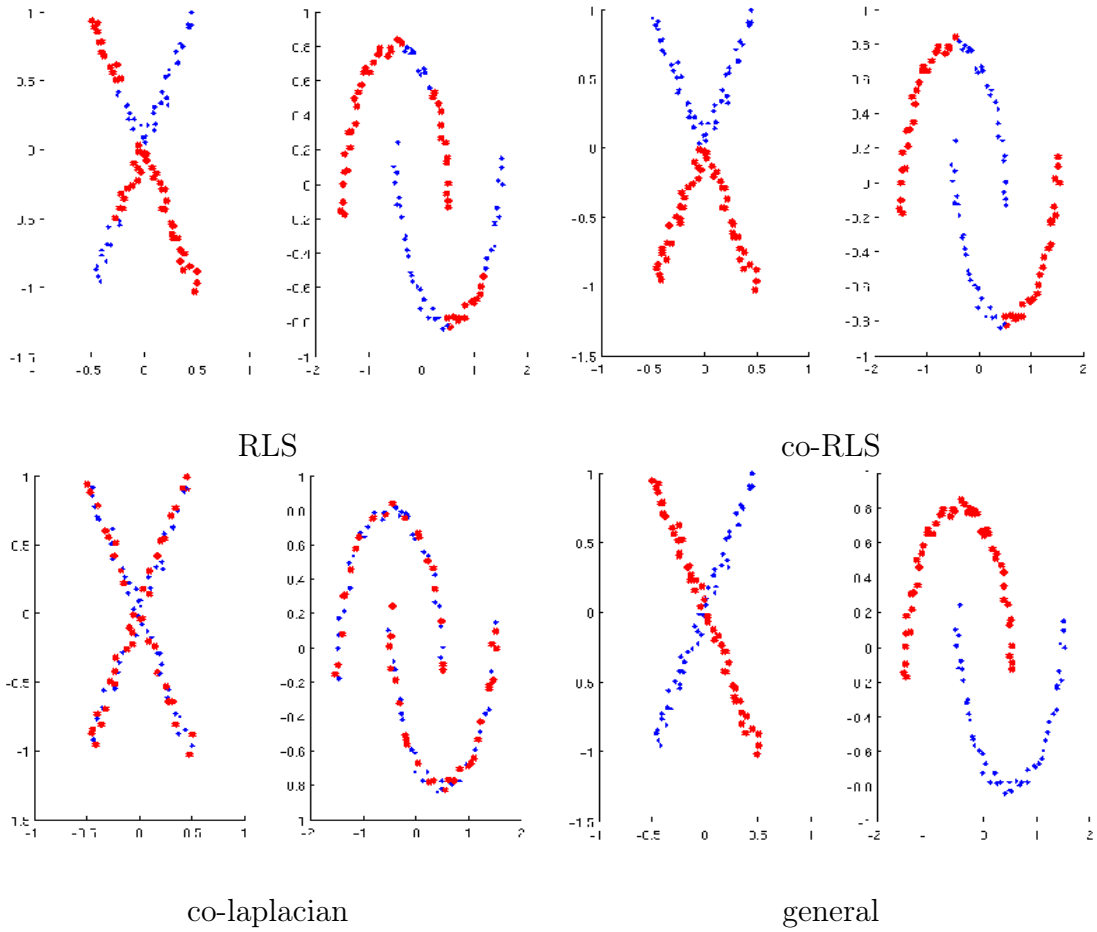


Figure 8.5: One cross - two moons. For each algorithm (RLS, co-RLS, co-laplacian, general), the classes learned on the objects projected on each view. Here only the general algorithm achieves exact classification

5.1 Previous work

The usual idea in supervised classification is to use cross-validation techniques (classically 10-fold). Although widely used, due to an easy implementation, it is generally admitted that cross-validation suffers from a lack of theory and understanding. This lack tends to be filled in. We refer for instance to Celisse (2008) for recent work on the theoretical analysis of cross-validation procedures. Another widely spread idea is to use Bayesian parameter selection (see Gold et al. (2005)). Standard RKHS hyper-tuning parameters may also been applied to the kernel underlying the general objective function.

Another approach comes from unsupervised learning theory (see Ben-David et al. (2006), Ben-David and von Luxburg (2008)). The idea is to work on stability of the clusterings output by the algorithms. This new issue makes use of recent improvements on statistical theory, coming from the analysis of empirical processes, Talagrand's work and so-called "small ball estimates" (see Li and Linde (1999),Berthet and Shi (2001)). Note that some refined bounds involving margin conditions, now usual in classification (Blanchard et al. (2003),Blanchard et al. (2008)), begin to be applied successfully (see the results of Sindhwani and Rosenberg (2008)). In Koltchinskii (2006), the author introduces and develops a stability-based parameter selection procedures. Based on this work, we promote a stability-based parameter selection for our setting. As mentioned earlier in this work, comparison of different parameter selection methods is not the aim of this work.

5.2 Theoretical selection procedure

Let $\mathcal{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure, and P the true measure. Thus $\mathcal{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $\mathbb{P}f = \mathbb{E}(f(X))$. For a general class $\mathcal{F}$ of functions, and probability measure Q, we define $\mathcal{F}_Q(\varepsilon) = \{f \in \mathcal{F}; Qf - \inf Qf \leqslant \varepsilon\}$ and then introduce the true ε-optimal ball $\mathcal{F}(\varepsilon) = \mathcal{F}_{\mathbb{P}}(\varepsilon)$, and the empirical ε-optimal ball $\mathcal{F}_n(\varepsilon) = \mathcal{F}_{\mathbb{P}_n}(\varepsilon)$, or balls around the Empirical Risk Minimizer (ERM) and True Risk Minimizer (TRM). For a general class $\mathcal{F}$ of functions, we now assume that we have $T : \mathcal{F}^2 \to \mathbb{R}^+$ such that $\forall f, g \in F \quad \mathbb{V}(f-g) \leqslant T^2(f,g)$, and then introduce the two objects: $\Delta_n(\varepsilon) = \sup_{f_1,f_2\in\mathcal{F}(\varepsilon)} |P_n - P|(f_1 - f_2)$ and $D_{\mathcal{F}}(\varepsilon) = \sup_{f,g\in\mathcal{F}(\varepsilon)} T(f,g)$. We refer to the first one as a L_1, P-diameter and the second one as a L_2, P-diameter. Lemma 8.4 in Koltchinskii (2006) says that for large enough radii, the empirical and true quasi-optimal balls around the ERM and TRM are included in each other, or put differently, that true quasi-optimal balls can be estimated by empirical quasi-optimal balls:

Lemma 8.4 (Koltchinskii) *$\forall \varepsilon > 0$, $\forall \lambda < 1$, set*

$$B_n(\varepsilon,\lambda) = 2\frac{\Delta_n(\varepsilon)}{\lambda} + \frac{\log(\varepsilon^{-1})}{\lambda n} + \frac{2}{\lambda}\sqrt{\frac{2\log(\varepsilon^{-1})}{n}[D^2_{\mathcal{F}}(\varepsilon) + 2\Delta_n(\varepsilon)]},$$

$$\textit{and } r_n(\varepsilon,\lambda) = \inf\{\alpha \in [0,1]; \sup_{j\in\mathbb{Z};1\geqslant\lambda^j\geqslant\alpha} B_n(\varepsilon,\lambda^j) \leqslant \lambda\}$$

Then with probability larger than $1 - \left(2 + \frac{\ln(r_n(\varepsilon,\lambda))}{\ln(\lambda)}\right)\varepsilon$,

$$\forall r \geqslant r_n(\varepsilon,\lambda) \quad \mathcal{F}(r) \subset \mathcal{F}_n(3r/2) \text{ and } \mathcal{F}_n(r) \subset \mathcal{F}(2r)$$

Note that in the general case, if the radii are too small, then such inclusions no longer hold, and the intersection may even be empty.

Having such inclusions means that we can stay around the Erm and yet still be close to the "true" minimizer of our functional, which is one way to see stability. Using this lemma on stability, we will simply select the parameter θ inducing the bigger range of quasi-optimal sets controlled around the ERM, which is a notion of stability, i.e. for a given radius ε of the true penalized ball, we want to minimize the critical radius r_n w.r.t. θ. A side motivating *intuition* is that having good stability allows for easy discovery of the minimizer f^*.

5.3 Empirical selection procedure

The two terms $\Delta_n(\varepsilon)$ and $D_{\mathcal{F}}(\varepsilon)$ of lemma 8.4 involve the true unknown measure and thus have to be estimated. We now propose an empirical version of this lemma. Fortunately, using an empirical estimation of the $r_n(\varepsilon,\lambda)$ is possible thanks to the Theorem 3, page 18, in Koltchinskii (2006), leading to a full data-dependent quantity. Let $\widehat{\Delta}_n(\varepsilon) = R_n(\mathcal{F}_n(\varepsilon))$ and $\widehat{D}_{\mathcal{F}_n}(\varepsilon) = \sup_{f,g\in\mathcal{F}_n(\varepsilon)} T_n(f,g)$, with T_n^2 bounding the empirical variance $\mathbb{V}_n$. The empirical version of the quantities $B_n(\varepsilon,\lambda)$ and $r_n(\varepsilon,\lambda)$ given by Koltchinskii (2006) are:

$$\widehat{r}_n(\varepsilon,\lambda) = \inf\left\{\alpha \in [0,1]; \sup_{j\in\mathbb{Z};1\geqslant\lambda^j\geqslant\alpha} \widehat{B}_n(\varepsilon,\lambda^j) \leqslant \lambda^3\right\},$$

where

$$\widehat{B}_n(\varepsilon,\lambda) = \frac{2c\widehat{\Delta}_n(c'\varepsilon)}{\lambda} + 2\widehat{D}_{\mathcal{F}_n}(c'\varepsilon)\sqrt{\frac{\log(\varepsilon^{-1})}{\lambda^2 n}} + \frac{\log(\varepsilon^{-1})}{\lambda n}$$

and $c, c' \geqslant 1$ are universal constants.

Application to semi-supervised multiview classification. We now propose to apply this result to semi-supervised multiview classification. We identify the classes $\widehat{\mathcal{F}}_{\theta,n}$ to be $\mathcal{J}(\theta,r)$, and estimate $R_n(\widehat{\mathcal{F}}_{\theta,n}(\varepsilon))$ and $\widehat{D}_{\widehat{\mathcal{F}}_{n,\theta}}(\varepsilon)$ for each parameter θ. Note the dependency w.r.t. $\theta = (\alpha,\lambda,\gamma,\mathcal{C})$ in the definition. Thus we need to bound the Rademacher complexity of $\mathcal{J}(r)$ and its L_2, P_n-Diameter. An analysis of the proof of Theorem 8.1 shows that changing $\mathcal{J} = \mathcal{J}(1)$ for $\mathcal{J}(r)$ affects the Rademacher bound with a factor $\sqrt{r}$, leading to a bound $\frac{2b(\theta)\sqrt{r}}{lV}$ for the first term. Following the same analysis as for the L_1-diameter (or Rademacher complexity), the next theorem gives us the second bound we need.

Theorem 8.3 (*Empirical local* L_2 *diameter*) *Under assumption A1, then the empirical local* L_2 *diameter of the class of decision functions is upper bounded by*

$$\widehat{D}_{\mathcal{J}}(r) \leqslant \frac{2d\sqrt{r}}{\sqrt{l}V}$$

where d^2 *is the largest eigenvalue of* $(B - J_2^T(I+M)^{-1}J_2)\tilde{\lambda}^{-1}\Pi(\underline{K_L^T})^T$*, with* $J_2 = \sqrt{C}\delta\tilde{\lambda}^{-1}B^T$*.*

Note that we here have a dependency with $\sqrt{l}$ instead of the l for the Rademacher bound. We have stated this theorem using the already defined matrices of Theorem 8.1. The proof consists of three steps. First reducing the supremum problem to a quadratic minimization problem, then use the change of matrix from N to T as in the proof of Theorem 8.1, and then rewrite the final solution in terms of the original data and kernel matrices.

Proof: By definition, we have

$$\widehat{D}_{\mathcal{J}}(r) = \sup_{\varphi_1,\varphi_2\in\mathcal{J}(r)} \mathbb{V}_l(\varphi_1-\varphi_2)^{1/2} = \sup_{\varphi_1,\varphi_2\in\mathcal{J}(r)} (\mathbb{P}_l((\varphi_1-\varphi_2)^2))^{1/2},$$

which can be upper-bounded by $\left[\frac{4}{l}\sup_{\varphi\in\mathcal{J}(r)}\sum_{i=1}^{l}\varphi(x_i)^2\right]^{1/2}$.

Since $\varphi = \frac{1}{V}\sum_{v=1}^{V} f^{(v)}$, and $f^{(v)}(x_i) = K_i^{(v)}\alpha^{(v)}$, where $K_i^{(v)}$ is the ith row of the kernel matrix $K^{(v)}$, we can upper bound $\widehat{D}_{\mathcal{J}}(r)^2$ by $\frac{4}{V^2 l}\underline{\alpha}^T D\underline{\alpha}$ where $D\in\mathbb{R}^{nV\times nV}$ is the matrix with block (v,w) equal to $\sum_{i=1}^{l} K_i^{(v)T}K_i^{(w)}$. Thus we want to solve the following problem, where D is symmetric: $\sup_{\underline{\alpha};\underline{\alpha}^T N\underline{\alpha}\leqslant r}\underline{\alpha}^T D\underline{\alpha}$. We use the following lemma:

Lemma 8.5 *When* M *is symmetric positive definite and* Q *is symmetric positive semi-definite, the quadratic problem :*

$$\sup_{a;a^T Ma\leqslant r} a^T Qa$$

admits as solution λr*, where* λ *is the highest eigenvalue of* $M^{-1}Q$*.*

One straightforward proof uses duality as in Lemma 8.2.

As before, we can not apply directly this result since the matrix N may not be invertible. Introducing instead the P, Σ matrices and a variable, of the proof of Theorem 8.1, the solution of our problem is λr, where λ is the highest eigenvalue of the matrix $T^{-1}\overline{P}^T D\overline{P}$.

Now, we can see that D is just $\underline{K}e\underline{K}^T$ with $e\in\mathbb{R}^{n\times n}$ being the projection matrix with diagonal blocks Id_l and 0_u. We use again the Sherman-Morrison-Woodbury formula to rewrite T^{-1} with the matrices A and U. Moreover, the eigenvalues of $T^{-1}\overline{P}^T D\overline{P}$ and $\overline{P}T^{-1}\overline{P}^T D$ are the same, since each P is a changing base matrix.

Thus, we compute the following term:

$$\begin{aligned}\overline{P}A^{-1}\overline{P}^T D &= \overline{PP}^T B\overline{P}\tilde{\lambda}^{-1}\overline{\Sigma}^{-1}\overline{P^T}\underline{K}e\underline{K}^T \\ &= B\overline{P}\tilde{\lambda}^{-1}Id_M\overline{P^T}\Pi(\underline{K_L^T})^T,\end{aligned}$$

where with our notation, $\overline{P}\tilde{\lambda}^{-1}Id_M\overline{P^T}$ reduces to $\tilde{\lambda}^{-1} \in \mathbb{R}^{nV\times nV}$. A similar computation yields the second term: $J_2^T(I+M)^{-1}J_2\tilde{\lambda}^{-1}\Pi(\underline{K_L^T})^T$. □

Eventually, for parameter selection, each θ leads to a radius $r_n^\theta(\varepsilon,\lambda) \geqslant \widehat{r}_n^\theta(\varepsilon,\lambda)$ defined likewise, but with the upper bound on $\widehat{\Delta}_n(c'\varepsilon)$ instead. For maximal stability, we propose to select the largest range of values for which the lemma still holds, i.e. minimize this quantity with θ. This leads to the selection procedure summed up below, where each term is explicit. Note that the supremum and infimum are computable.

- Fix a probability threshold with $\varepsilon > 0$,$\lambda < 1$.
- Compute $r(\theta,n,l,\varepsilon,\lambda)$, defined by:

$$\inf\left\{\alpha\in[0,1];\ \sup_{j\in\mathbb{Z};1\geqslant\lambda^j\geqslant\alpha}\tilde{B}_{n,l}(\theta,\varepsilon,\lambda^j)\leqslant\lambda^3\right\},$$

  where the term $\tilde{B}_{n,l}(\varepsilon,\lambda)$ is:

$$\frac{2cb(\theta)\sqrt{c'\varepsilon}}{lV\lambda}+\frac{4d(\theta)\sqrt{c'\varepsilon}}{\sqrt{l}V}\sqrt{\frac{\log(\varepsilon^{-1})}{\lambda^2 n}}+\frac{\log(\varepsilon^{-1})}{\lambda n}$$

- Output: $\theta^* = \arg\min_{\theta\in\Theta} r(\theta,n,l,\varepsilon,\lambda)$

Parameter selection procedure for multiple-view semi-supervised learning.

Conclusion

In this chapter, we have combined different aspects of semi-supervised and multiview learning into one algorithm for which we provide explicit error bounds. Moreover, we have performed the analysis for the full multiview learning problem which is a meaningful improvement to the two-view problem. Besides, we have combined stability concepts from the statistical and clustering communities to propose a new stability-based parameter selection, which benefits from recent theoretical developments. Additionally, we provide an explicit stability (L_2-diameter) bound for each parameter, which has not been investigated so far.

Part III

Towards the Real World(?): Modeling and Planning.

In this last part, we gather the world of bandits together with the batch world, hopefully for the better, in order to address some reinforcement learning problems.

Markov Decision Processes, Value function and Bellman operator. In a nutshell - see Sigaud and Buffet (2008) for an introductory book written in french on this topic - the standard reinforcement learning (RL) framework (Bertsekas and Tsitsiklis, 1996, Sutton and Barto, 1998) considers a learning agent that interacts with a stochastic environment and this interaction is modeled as a discrete-time discounted Markov decision process (MDP). A discounted MDP is a tuple $\mathcal{M} = \langle \mathcal{X}, \mathcal{A}, r, P, \gamma \rangle$ where the state space $\mathcal{X}$ is a bounded closed subset of a Euclidean space, $\mathcal{A}$ is a finite ($|\mathcal{A}| < \infty$) action space, the reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is uniformly bounded by $R_{\max}$, the transition kernel P is such that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $P(\cdot|x, a)$ is a distribution over $\mathcal{X}$, and $\gamma \in (0, 1)$ is a discount factor.

The optimal value function V^* is the unique fixed-point of the so-called optimal Bellman operator $\mathcal{T}^* : \mathcal{B}(\mathcal{X}; V_{\max} \stackrel{\text{def}}{=} \frac{R_{\max}}{1-\gamma}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by

$$(\mathcal{T}^* V)(x) = \max_{a \in \mathcal{A}} \left[r(x, a) + \gamma \int_{\mathcal{X}} p(dy|x, a) V(y) \right]. \tag{8.2}$$

where $B(\mathcal{X}; L)$ is the space of measurable functions with domain $\mathcal{X}$ bounded by $L < \infty$.

A deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$ is a mapping from states to actions. The general goal of reinforcement learning is to learn a policy that enables to receive maximal discounted cumulative reward from any given initial state. One possible way do to that is to learn the optimal value function V^*, since the greedy policy corresponding to V^* is actually the optimal one in that sense. In the sequel, we focus on this approach.

One traditional idea in order to address this problem is to apply a *policy iteration* algorithm: Starting from some initial policy, we repeat two steps where we first estimate the value function V^π of the current policy π (see definition in Section 2 of chapter 9) using some sampling scheme and then improve the current policy based on this estimation.

The difficulty to estimate V^π typically depends on whether the space $\mathcal{X}$ is finite or not, and whether we are forced to sample the immediate reward and the next state only by following a single trajectory from the initial state or whether we can resort to a *generative model* that enables to sample the immediate reward and the next state from any desired state, thus not necessarily the current one. The first case corresponds to chapter 10 while the second corresponds to chapter 9.

Another way to address this problem is to apply the optimism in face of uncertainty principle: Starting from some initial policy, we repeat two steps where we first compute, based on the current samples, a set of plausible optimistic models of MDPs together with the corresponding optimistic policy - this is possible since all transitions and reward are known in the optimistic models - and then get new samples by running the optimistic policy. This is the point of view used in chapter 11 and introduced in Jaksch et al. (2010).

Of course there are plenty of other methods, see (Bertsekas and Tsitsiklis, 1996, Sutton and Barto, 1998) or more recently Szepesvári (2010) for further details.

Contributions. The following three contributions all make use of concepts coming from statistical learning theory as well as from bandit theory, and address a specific reinforcement learning issue.

In chapter 9, we analyze an algorithm called Bellman residual minimization that is a natural algorithm in the setting of discounted MDP where we are allowed to resort to a generative model, i.e. we can sample at any time one action from any possible state, as opposed to the setting where we can only sample one action from the current state.

In chapter 10, we analyze a version of an algorithm called Least-squares Temporal Difference, where we make use of random projections as developed in chapter 6, in order to benefit from dimension reduction. This algorithm is designed for discounted MDP in the case when we do not have a generative model and thus we are forced to sample actions according to the current state, following one trajectory. Interestingly, from a statistical learning theory point of view, the value estimation problem corresponding to this algorithm can be seen as a regression problem with Markov design, where the target value function can not be sampled directly, but instead is defined as the fix point of the Bellman operator that we need to estimate.

Finally in chapter 11, we pave the way towards addressing the important problem of selecting a model of states for reinforcement learning. Indeed, in practice it may be difficult to define a good notion of states, and thus there may be different possible modelizations. We build our analysis on top of the UCRL2 algorithm designed for non-discounted MDPs, and consider a setting that is philosophically related to the chapter 4 that targets the challenging question of adaptive bandits.

Chapter 9

Finite-Sample Analysis of the Bellman Residual Minimization algorithm.

We consider the Bellman residual minimization approach for solving discounted Markov decision problems, where we assume that a generative model of the dynamics and rewards is available. At each policy iteration step, an approximation of the value function for the current policy is obtained by minimizing an empirical Bellman residual defined on a set of n states drawn i.i.d. from a distribution μ, the immediate rewards, and the next states sampled from the model. Our main result is a generalization bound for the Bellman residual in linear approximation spaces. In particular, we prove that the empirical Bellman residual approaches the true (quadratic) Bellman residual in μ-norm with a rate of order $O(1/\sqrt{n})$. This result implies that minimizing the empirical residual is indeed a sound approach for the minimization of the true Bellman residual which guarantees a good approximation of the value function for each policy. Finally, we derive performance bounds for the resulting approximate policy iteration algorithm in terms of the number of samples n and a measure of how well the function space is able to approximate the sequence of value functions.

The work presented in this chapter is a joint work with *Alessandro Lazaric* and *Mohammad Ghavamzadeh* and has been published in the proceedings of the *Asian Conference on Machine Learning (ACML 2010)*, see Maillard et al. (2010).

Contents

1 Introduction

In this paper we consider the problem of solving a Markov decision problem (MDP) (Bertsekas and Shreve, 1978, Puterman, 1994) by means of an approximate policy iteration algorithm (Bertsekas and Tsitsiklis, 1996, Si et al., 2004, Powell, 2007) with a linear approximation space $\mathcal{F}$. In particular, we focus on the Bellman residual minimization approach (Schweitzer and Seidmann, 1985, Baird, 1995, Munos, 2003, Lagoudakis and Parr, 2003, Scherrer, 2010) when a generative model is available, that is, for any state-action pair it is possible to obtain the immediate reward and an independent sample of the next state drawn from the transition distribution.

More in details, at each iteration k, in order to evaluate the current policy π_k, we build an approximation $V_k \in \mathcal{F}$ of the value function V^{π_k} by solving an empirical Bellman residual minimization problem: $V_k = \arg\min_{f\in\mathcal{F}} \mathcal{B}_n(f)$, where $\mathcal{B}_n(f)$ is the empirical Bellman residual. The specific definition of $\mathcal{B}_n$ is critical since, as observed in several previous works (see e.g., Sutton and Barto 1998, Lagoudakis and Parr 2003, Antos et al. 2008), the squared temporal difference between successive states (e.g., states obtained following a single trajectory), gives rise to a biased estimate of the (true) Bellman residual $\mathcal{B}(f) = ||f - \mathcal{T}^\pi f||_\mu^2$. In this paper, in order to build an unbiased estimate of $\mathcal{B}(f)$ we take advantage of the generative model and build $\mathcal{B}_n$ on n states drawn i.i.d. from a given distribution μ, as well as the immediate rewards and two next states independently sampled from the generative model (i.e., the double-sampling technique suggested in Sutton and Barto 1998, p. 220).

Motivation. The idea of minimizing the Bellman residual is natural (see e.g., Schweitzer and Seidmann 1985, Baird 1995) and it is based on the property that for any policy π the value function V^π has a zero residual, i.e., $\mathcal{B}(V^\pi) = 0$. As a result, it is reasonable to expect that the minimization of the Bellman residual $\mathcal{B}(f)$ in a given function space $\mathcal{F}$ leads to a function which is close to the value function. Williams and Baird (1994) and Munos (2007) proved that indeed the residual $||\mathcal{T}^\pi f - f||$ (in sup-norm and L_p-norms, respectively) of a function f is related to its distance (in the same norm) to the value function V^π, $||V^\pi - f||$. Thus, minimizing the Bellman residual leads to a small approximation error. However, those results concern the (true) Bellman residual $\mathcal{B}(f)$ but not its empirical estimate $\mathcal{B}_n(f)$, which is the quantity that is actually minimized by real algorithms.

Although it is often believed that the minimization of the empirical residual $\mathcal{B}_n(f)$ is "approximately" equivalent to minimizing the (true) residual $B(f)$, no such result is available in the literature so far. The closest work in this direction is by Antos et al. (2008), who provides a finite-sample analysis of a variant of the Bellman-residual minimization, called Modified Bellman residual, which reduces to Least Squares Temporal Differences (LSTD) in the case of linear function spaces. A finite sample analysis of LSTD is also reported in Lazaric et al. (2010c), and a regularized version of those algorithms is described in Farahmand et al. (2008). However, these works analyze algorithms that are related but different from the empirical Bellman residual minimization considered here.

Contribution. Our main contribution in this paper is to address this question: does

minimizing the empirical Bellman residual $\mathcal{B}_n$ implies that we also minimize the true Bellman residual at all states w.r.t. a distribution μ? In other terms, is it possible to control the true Bellman residual $\mathcal{B}(f)$ in terms of the empirical Bellman residual $\mathcal{B}_n(f)$?

We show that the answer to those questions is actually not obvious but is positive. It is not obvious because we show that the usual generalization results for regression cannot be trivially adopted in bounding the difference between the true Bellman residual and its empirical counterpart. In fact, in Bellman residual minimization we are not trying to minimize an empirical distance to a given target function, but we are directly searching for an approximate fixed-point (in $\mathcal{F}$) of an empirical version of the Bellman operator $\mathcal{T}^\pi$. As a result, it might be possible that a function with very low empirical residual (even possibly zero) at the sampled states has a large (true) Bellman residual at other states and even at the same states. However, we show that this problem does not occurs when the empirical Bellman residual minimizer belongs to a set of controlled sized (e.g. measured in terms of the norm of its parameter). More precisely, we show that for functions $f_\alpha \in \mathcal{F}$ with bounded parameter $||\alpha||$, the difference between $\mathcal{B}(f_\alpha)$ and $\mathcal{B}_n(f_\alpha)$ decreases as the number of samples n increases. Then, we prove that when the number of samples n is large enough, the norm $||\widehat{\alpha}||$ of the empirical Bellman residual minimizer $f_{\widehat{\alpha}} = \arg\min_{f_\alpha \in \mathcal{F}} \mathcal{B}_n(f)$ is indeed upper-bounded, provided that the set of features defining the linear space $\mathcal{F}$ are linearly independent under the distribution μ. Thus we deduce that the Bellman residual $\mathcal{B}(f_{\widehat{\alpha}})$ of the empirical Bellman minimizer $f_{\widehat{\alpha}}$ is bounded by the empirical Bellman residual $\mathcal{B}_n(f_{\widehat{\alpha}})$ plus an estimation error term of order $O(1/\sqrt{n})$. In other terms, we provide a generalization result for the Bellman residual in linear approximation spaces. This result implies that minimizing the empirical residual is indeed a sound approach for deriving a good approximation of the value function for each policy.

The paper is organized as follows. In Section 2 we introduce the notation. Section 3 reports the main contribution of this paper, that is the finite-sample analysis of Bellman residual minimization for policy evaluation. Finally, in Section 4 we extend the policy evaluation result to the whole policy iteration algorithm.

2 Preliminaries

In this section, we introduce the main notations used in the paper. For a measurable space with domain $\mathcal{X}$, we let $\mathcal{S}(\mathcal{X})$ and $\mathcal{B}(\mathcal{X}; L)$ denote the set of probability measures over $\mathcal{X}$ and the space of bounded measurable functions with domain $\mathcal{X}$ and bound $0 < L < \infty$, respectively. For a measure $\mu \in \mathcal{S}(\mathcal{X})$ and a measurable function $f : \mathcal{X} \to \mathbb{R}$, we define the $\ell_2(\mu)$-norm of f as $||f||_\mu^2 = \int f(x)^2 \mu(dx)$, the supremum norm of f as $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Moreover, for a vector $u \in \mathbb{R}^d$, we write its ℓ_2-norm as $||u||^2 = \sum_{i=1}^d u_i^2$.

We consider the standard reinforcement learning (RL) framework (Bertsekas and Tsitsiklis, 1996, Sutton and Barto, 1998) in which a learning agent interacts with a stochastic environment and this interaction is modeled as a discrete-time discounted Markov decision

process (MDP). A discounted MDP is a tuple $\mathcal{M} = \langle \mathcal{X}, \mathcal{A}, r, P, \gamma \rangle$ where the state space $\mathcal{X}$ is a bounded closed subset of a Euclidean space, $\mathcal{A}$ is a finite ($|\mathcal{A}| < \infty$) action space, the reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is uniformly bounded by $R_{\max}$, the transition kernel P is such that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $P(\cdot|x, a)$ is a distribution over $\mathcal{X}$, and $\gamma \in (0, 1)$ is a discount factor. A deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$ is a mapping from states to actions. Under a policy π, the MDP $\mathcal{M}$ is reduced to a Markov chain $\mathcal{M}^\pi = \langle \mathcal{X}, R^\pi, P^\pi, \gamma \rangle$ with reward $R^\pi(x) = r(x, \pi(x))$ and transition kernel $P^\pi(\cdot|x) = P(\cdot|x, \pi(x))$.

Value functions. The value function of a policy π, V^π, is the unique fixed-point of the Bellman operator $\mathcal{T}^\pi : \mathcal{B}(\mathcal{X}; V_{\max} = \frac{R_{\max}}{1-\gamma}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by

$$(\mathcal{T}^\pi V)(x) = R^\pi(x) + \gamma \int_{\mathcal{X}} P^\pi(dy|x)V(y). \tag{9.1}$$

We also define the optimal value function V^* as the unique fixed-point of the optimal Bellman operator $\mathcal{T}^* : \mathcal{B}(\mathcal{X}; V_{\max} = \frac{R_{\max}}{1-\gamma}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by

$$(\mathcal{T}^* V)(x) = \max_{a \in \mathcal{A}} \left[r(x, a) + \gamma \int_{\mathcal{X}} p(dy|x, a)V(y) \right]. \tag{9.2}$$

Approximation space. We consider a linear function space $\mathcal{F}$ defined as the span of d basis functions $\varphi_i : \mathcal{X} \mapsto \mathbb{R}$, $i = 1, \ldots, d$, i.e.,

$$\mathcal{F} = \{ f_\alpha(\cdot) = \varphi(\cdot)^\top \alpha, \alpha \in \mathbb{R}^d \},$$

where $\varphi(\cdot) = \big(\varphi_1(\cdot), \ldots, \varphi_d(\cdot)\big)^\top$ is the feature vector. We define the Gram matrix $G \in \mathbb{R}^{d \times d}$ with respect to a distribution $\mu \in \mathcal{S}(\mathcal{X})$ as

$$G_{ij} = \int_{\mathcal{X}} \varphi_i(x)\varphi_j(x)\mu(dx), \tag{9.3}$$

with $i, j = 1, \ldots, d$. Finally, we write $L_{\max} = \sup_{x \in \mathcal{X}} ||\varphi(x)||$ and assume that $L_{\max} < \infty$.

3 Bellman Residual Minimization for Policy Evaluation

In this section, we consider the Bellman Residual Minimization (BRM) algorithm for the evaluation of a fixed policy π, using the double sampling technique (see e.g., Sutton and Barto 1998). We assume that a generative model of the MDP is available, and that for each state x and action a a call to the generative model returns the reward $r(x, a)$ and two independent samples drawn from the distribution $P(\cdot|x, a)$.

3.1 The Empirical Bellman Residual Solution

We build a dataset $\mathcal{D} = \{(X_i, R_i, Y_i, Y_i')_{1 \leqslant i \leqslant n}\}$ where for all $i = 1, \ldots, n$, we sample a state $X_i \overset{iid}{\sim} \mu$ and make a call to the generative model to obtain the reward $R_i = r(X_i, \pi(X_i))$ and two independent next-state samples Y_i and Y_i' drawn from $P^\pi(\cdot|X_i)$. The **empirical Bellman residual** (EBR) is defined for any $f \in \mathcal{F}$ as

$$\mathcal{B}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \left[f(X_i) - \gamma f(Y_i) - R_i \right] \left[f(X_i) - \gamma f(Y_i') - R_i \right]. \tag{9.4}$$

The EBR minimizer $f_{\widehat{\alpha}}$ is defined, whenever it exists, as the minimizer of $\mathcal{B}_n(f_\alpha)$ in $\mathcal{F}$:

$$f_{\widehat{\alpha}} = \arg\min_{f_\alpha \in \mathcal{F}} \mathcal{B}_n(f_\alpha), \tag{9.5}$$

and $\widehat{\alpha}$ is the parameter of the EBR minimizer. Using matrix notations, by defining the $n \times d$-matrices Ψ and Ψ' as $\Psi_{ij} = \varphi_j(X_i) - \gamma \varphi_j(Y_i)$ and $\Psi'_{ij} = \varphi_j(X_i) - \gamma \varphi_j(Y_i')$, $\mathcal{B}_n(f_\alpha)$ may be written as

$$\mathcal{B}_n(f_\alpha) = \frac{1}{n} \left[\alpha^\top \Psi^\top \Psi' \alpha - R^\top (\Psi + \Psi') \alpha + R^\top R \right],$$

where $R \in \mathbb{R}^n$ is the vector of components R_i. Thus, by defining the $d \times d$ empirical Gram matrix $A = \frac{1}{n}(\Psi^\top \Psi' + \Psi'^\top \Psi)$, the d-vector $b = \frac{1}{n}(\Psi + \Psi')^\top R$, and the constant $c = \frac{1}{n} R^\top R$, we have

$$\mathcal{B}_n(f_\alpha) = \frac{1}{2} \alpha^\top A \alpha - b^\top \alpha + c. \tag{9.6}$$

Using this notation, the gradient of $\mathcal{B}_n$ is $\nabla_\alpha \mathcal{B}_n(f_\alpha) = A\alpha - b$, thus whenever the EBR minimizer exists, its parameter $\widehat{\alpha}$ is the solution to the linear system $A\alpha = b$.

Although the empirical Bellman residual $\mathcal{B}_n(f_\alpha)$ is a quadratic function of α, with A a symmetric matrix, A may not be definite positive. A may even possess negative eigenvalues, thus $\mathcal{B}_n(f_\alpha)$ may not have any minimizer. However we will see in the next section that when n is large enough then the EBR minimizer exists and is unique.

3.2 Finite-Sample Analysis

Defining $\mathcal{B}(f) = ||f - \mathcal{T}^\pi f||_\mu^2$ the true squared Bellman residual in μ-norm, we have the property that for any f, $\mathcal{B}_n(f)$ is an unbiased estimate of $\mathcal{B}(f)$. In fact,

$$\mathbb{E}_{Y_i, Y_i' \overset{iid}{\sim} P^\pi(\cdot|X_i)} \Big[\left[f(X_i) - \gamma f(Y_i) - R_i \right] \left[f(X_i) - \gamma f(Y_i') - R_i \right] | X_i \Big] = \left[f(X_i) - \mathcal{T}^\pi f(X_i) \right]^2,$$

thus, since $X_i \overset{iid}{\sim} \mu$, it follows that $\mathbb{E}_{\mathcal{D}}[\mathcal{B}_n(f)] = \mathcal{B}(f)$.

The main issue is to show that by minimizing the empirical Bellman residual $\mathcal{B}_n$, we actually obtain a function $f_{\widehat{\alpha}}$ whose (true) residual $f_{\widehat{\alpha}} - \mathcal{T}^\pi f_{\widehat{\alpha}}$ is small at the states $(X_1, \ldots, X_n)$ and at other states measured by μ (i.e., it has a small $\mathcal{B}$). This property would hold if we could have a generalization result for the Bellman residual, like in the regression setting.

In regression, generalization bounds for spaces bounded in sup-norm are applied to the result of the truncation (at a threshold which depends on a sup-norm of the target function) of the empirical risk minimizer (Györfi et al., 2002). However, this approach does not work for BRM, because the truncation $\bar{f}_{\widehat{\alpha}}$ of the EBR minimizer $f_{\widehat{\alpha}}$ may amplify the residual (i.e., $\mathcal{B}(\bar{f}_{\widehat{\alpha}})$ may not be smaller than $\mathcal{B}(f_{\widehat{\alpha}})$). Thus, we follow another direction by considering spaces of functions $\mathcal{F}(C) \subset \mathcal{F}$ with bounded parameter: $\mathcal{F}(C) = \{f_\alpha \in \mathcal{F}, ||\alpha|| \leqslant C\}$, and provide a generalization bound for Bellman residual for functions $f_\alpha \in \mathcal{F}(C)$ (the proof is in Section 6).

Lemma 9.1 *For any $\delta > 0$, we have that with probability at least $1 - \delta$,*

$$\sup_{f_\alpha \in \mathcal{F}(C)} |\mathcal{B}(f_\alpha) - \mathcal{B}_n(f_\alpha)| \leqslant c_1 \sqrt{\frac{2d\log(2) + 6\log(8/\delta)}{n}},$$

where $c_1 = 96\sqrt{2}[C(1+\gamma)L_{\max} + R_{\max}]^2$.

Unfortunately, this result cannot be immediately applied to the EBR minimizer $f_{\widehat{\alpha}}$ since we do not have a bound on the norm $||\widehat{\alpha}||$. In fact, when we solve the minimization problem (9.5), we do not have any control on the norm of the solution (if it exists) $||\widehat{\alpha}||$. For instance, if we consider the case in which two features φ_1 and φ_2 are identical, then $\alpha_1\varphi_1 + \alpha_2\varphi_2 = 0$ whenever $\alpha_1 = -\alpha_2$, thus $||\alpha||$ can be made arbitrarily large without changing the value of f_α simply by playing on the values of α_1 and α_2. In order to avoid such degenerate situations, we introduce the following assumption on the linear independence of the features $(\varphi_i)_{1 \leqslant i \leqslant d}$ w.r.t. the distribution μ.

Assumption The smallest eigenvalue ν of the Gram matrix G (defined in (9.3)) is strictly positive, i.e., $\nu > 0$. [1]

We show in the following that Assumption 3.2 is a sufficient condition to derive a bound on the norm $||\widehat{\alpha}||$ for any $\widehat{\alpha}$ solution of the EBR minimization problem. Before moving to the analysis of the EBR minimizer with linear independent features, we first introduce some additional notation. Let $\mathcal{L}(f) = ||(I - \gamma P^\pi)f||_\mu^2$ be the quadratic part of $\mathcal{B}(f)$, and

$$\mathcal{L}_n(f) = \frac{1}{n}\sum_{i=1}^n \left[f(X_i) - \gamma f(Y_i)\right]\left[f(X_i) - \gamma f(Y_i')\right],$$

be its empirical version. Thus $\mathcal{L}_n(f_\alpha) = \frac{1}{2}\alpha^\top A\alpha$. Now, whenever the EBR minimizer $f_{\widehat{\alpha}}$ exists, since by definition $\widehat{\alpha}$ satisfies $A\widehat{\alpha} = b$, we can write

$$\mathcal{B}_n(f_\alpha) = \frac{1}{2}(\alpha - \widehat{\alpha})^\top A(\alpha - \widehat{\alpha}) - \frac{1}{2}\widehat{\alpha}^\top A\widehat{\alpha} + c = \mathcal{L}_n(f_{\alpha - \widehat{\alpha}}) - \mathcal{L}_n(f_{\widehat{\alpha}}) + c, \tag{9.7}$$

and deduce that $\mathcal{B}_n(f_{\widehat{\alpha}}) = c - \mathcal{L}_n(f_{\widehat{\alpha}}) = c - \frac{1}{2}b^\top\widehat{\alpha}$.

[1] Note that this condition implies the linear independence of the features in μ-norm.

Bounding $||\widehat{\alpha}||$. In order to deduce a bound on the parameter of the EBR minimizer $\widehat{\alpha}$, in the next three lemmas, we relate $||\alpha||$ to respectively $\mathcal{L}(f_\alpha)$ and $\mathcal{L}_n(f_\alpha)$. For that purpose, let us write

$$C^\pi(\mu) = (1-\gamma)||(I-\gamma P^\pi)^{-1}||_\mu,$$

which is related to the concentrability coefficient (see e.g., Antos et al. 2008) of the discounted future state distribution starting from μ and following policy π, i.e., $(1-\gamma)\mu(I-\gamma^\pi)^{-1}$ w.r.t. μ. Note that if the discounted future state distribution is not absolutely continuous w.r.t. μ, then $C^\pi(\mu) = \infty$.

Lemma 9.2 *Under Assumption 3.2, for any $\alpha \in \mathbb{R}^d$*

$$||\alpha||^2 \leqslant \frac{1}{\nu}||f_\alpha||_\mu^2 \leqslant \frac{C^\pi(\mu)^2}{\nu(1-\gamma)^2}\mathcal{L}(f_\alpha).$$

This indicates that the eigenvalues of the Gram matrix $\widetilde{G}$ defined by $\widetilde{G}_{ij} = \int_{\mathcal{X}} \psi_i \psi_j d\mu$, where $\psi_i = (I-\gamma P^\pi)\varphi_i$, are lower bounded by $\xi = \frac{\nu(1-\gamma)^2}{C^\pi(\mu)^2}$.

Proof: From the definition that ν is the smallest eigenvalue of G, we have $\alpha^\top\alpha \leqslant \frac{1}{\nu}\alpha^\top G\alpha = \frac{1}{\nu}||f_\alpha||_\mu^2$. Now since $(I-\gamma P^\pi)$ is an invertible operator (the eigenvalues of any stochastic kernel P^π have a modulus less than 1), we have $||f_\alpha||_\mu^2 \leqslant ||(I-\gamma P^\pi)^{-1}||_\mu^2||(I-\gamma P^\pi)f_\alpha||_\mu^2 = \left(\frac{C^\pi(\mu)}{1-\gamma}\right)^2\mathcal{L}(f_\alpha)$, and the lemma follows. □

This lemma provides a bound on $||\widehat{\alpha}||$ in terms of $\mathcal{L}(f_{\widehat{\alpha}})$. However $\mathcal{L}(f_{\widehat{\alpha}})$ is not known, and we would like to relate it to its empirical counterpart $\mathcal{L}_n(\widehat{\alpha})$. The next lemma (the proof is in Section 6) provides a generalization bound for $\mathcal{L}$, which enables to bound the difference between $\mathcal{L}$ and $\mathcal{L}_n$.

Lemma 9.3 *For any $\delta > 0$, we have that with probability at least $1-\delta$,*

$$\forall \alpha \in \mathbb{R}^d, \quad |\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| \leqslant c_2||\alpha||^2\sqrt{\frac{2d\log(2) + \log(4/\delta)}{n}},$$

where $c_2 = 96\sqrt{2}(1+\gamma)^2 L_{\max}^2$.

Combining Lemmas 9.2 and 9.3 we deduce that when n is large enough (as a function of ν and $C^\pi(\mu)$), then all the eigenvalues of the empirical Gram matrix A are strictly positive, and thus the EBR minimizer exists and is unique.

Lemma 9.4 *For any $\delta > 0$, whenever $n \geqslant n^\pi(\nu,\delta) = \frac{4c_2^2 C^\pi(\mu)^4}{\nu^2(1-\gamma)^4}(2d\log 2 + \log 4/\delta)$, with probability $1-\delta$ we have for all $\alpha \in \mathbb{R}^d$, $||\alpha||^2 \leqslant \frac{2}{\xi}\mathcal{L}_n(f_\alpha)$.*

We deduce that all the eigenvalues of the empirical Gram matrix A are strictly positive, and thus the EBR minimizer exists and is unique.

Proof: From Lemmas 9.2 and 9.3,

$$||\alpha||^2 \leqslant \frac{1}{\xi}\mathcal{L}(f_\alpha) \leqslant \frac{1}{\xi}\big(\mathcal{L}_n(f_\alpha) + c_2||\alpha||^2\sqrt{\frac{2d\log(2)+\log(4/\delta)}{n}}\big),$$

thus whenever $c_2\sqrt{\frac{2d\log(2)+\log(4/\delta)}{n}} \leqslant \frac{\xi}{2}$, i.e., $n \geqslant n^\pi(\nu,\delta)$, we have $||\alpha||^2 \leqslant \frac{2}{\xi}\mathcal{L}_n(f_\alpha)$. The claim about the eigenvalues of the empirical Gram matrix simply follows from the statement of the Lemma, the inequality $\alpha^\top\alpha \leqslant \frac{1}{\chi}\alpha^\top A\alpha$, where χ is the smallest eigenvalue of A, and the definition of $\mathcal{L}_n(f_\alpha) = \frac{1}{2}\alpha^\top A\alpha$. □

From this result we immediately deduce a bound on $||\widehat{\alpha}||$.

Corollary 9.1 *For any $\delta > 0$, whenever $n \geqslant n^\pi(\nu,\delta)$, with probability $1-\delta$ we have*

$$||\widehat{\alpha}|| \leqslant \frac{2}{\xi}(1+\gamma)L_{\max}R_{\max}.$$

Proof: From Lemma 9.4, using Cauchy-Schwarz's inequality, and recalling the definition of Ψ in Section 3.1

$$\begin{aligned}
||\widehat{\alpha}||^2 &\leqslant \frac{2}{\xi}\frac{1}{2}b^\top\widehat{\alpha} = \frac{1}{\xi}\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{i=1}^{n}R_i(\Psi_{i,j}+\Psi'_{i,j})\widehat{\alpha}_j\right)\\
&\leqslant \frac{1}{\xi}\frac{1}{n}\sum_{i=1}^{n}R_{\max}\Big(2\Big|\sum_{j=1}^{d}\widehat{\alpha}_j\varphi_j(X_i)\Big| + \gamma\Big|\sum_{j=1}^{d}\widehat{\alpha}_j\varphi_j(Y_i)\Big| + \gamma\Big|\sum_{j=1}^{d}\widehat{\alpha}_j\varphi_j(Y'_i)\Big|\Big)\\
&\leqslant \frac{2}{\xi}R_{\max}||\widehat{\alpha}||\sup_x||\varphi(x)||(1+\gamma)
\end{aligned}$$

from which the result follows. □

We now state our main result which bounds the Bellman residual of the EBR minimizer.

Theorem 9.1 (*Bellman residual of BRM*) *For any $\delta > 0$, whenever $n \geqslant n^\pi(\nu,\delta/2)$, with probability $1-\delta$ we have*

$$\mathcal{B}(f_{\widehat{\alpha}}) \leqslant \mathcal{B}_n(f_{\widehat{\alpha}}) + c_3\sqrt{\frac{2d\log(2)+6\log(8/\delta)}{n}},$$

where $c_3 = 96\sqrt{2}[\frac{2}{\xi}(1+\gamma)^2L_{\max}^2+1]^2R_{\max}^2$.

Proof: When $n \geqslant n^\pi(\nu,\delta)$, Corollary 9.1 states that $||\widehat{\alpha}|| \leqslant C$ is bounded and the results follows from a direct consequence of Lemma 9.1. □

Thus, the true residual $\mathcal{B}(f_{\widehat{\alpha}})$ of the EBR minimizer $f_{\widehat{\alpha}}$ is upper-bounded by the empirical residual $\mathcal{B}_n(f_{\widehat{\alpha}})$ plus an estimation error term, which is of order $O(1/\sqrt{n})$. We deduce that minimizing the empirical residual is indeed a sound method for deriving a function with small (true) Bellman residual $\mathcal{B}$.

Remark 1 The obtained estimation error term is of order $O(1/\sqrt{n})$, which is worse than the estimation error of order $O(\log n/n)$ deduced in linear regression with a quadratic loss (see e.g., Györfi et al. 2002). This is due to the fact that although $\mathcal{B}(f)$ is positive for any f, this is not the case for $\mathcal{B}_n(f)$, which may be negative (e.g., think of $\mathcal{B}_n(V^\pi)$ which is an unbiased estimate of $\mathcal{B}(V^\pi) = 0$). Thus the usual argument described in Györfi et al. (2002), where one would derive $\sqrt{\mathcal{B}(f)} \leqslant 2\sqrt{\mathcal{B}_n(f)} + O(1/\sqrt{n})$ does not directly apply here. One could also think of applying this argument to $\mathcal{L}_n$, since $\mathcal{L}_n$ is positive for sufficiently large n. However, this does not work either, since $\mathcal{L}_n$ is the sum of terms which are not individually positive, independently of the value of n. Therefore, it remains an open question to whether it is possible to obtain a bound of the form $\mathcal{B}(f_{\widehat{\alpha}}) \leqslant c\mathcal{B}_n(f_{\widehat{\alpha}}) + O(\log n/n)$ (with an additional multiplicative factor $c > 1$). This could be particularly interesting when $\mathcal{B}_n(f_{\widehat{\alpha}})$ is small.

Remark 2 The dependence to the dimension d of the function space $\mathcal{F}$ is of order $L_{\max}^4\sqrt{d}$. This is due to the fact that we cannot use truncation in this Bellman residual setting (see the first paragraph of Section 3.2), which would give us an order $L_{\max}^2\sqrt{d}$. We use instead a covering of the function space $\mathcal{F}(C)$ (see Theorem 9.2) with C (which itself depends on $L_{\max}$) being a bound on $||\widehat{\alpha}||$. This explains the additional $L_{\max}^2$ factor.

Remark 3 It is interesting to notice that although we derived Corollary 9.1 specifically for the case of Bellman residual minimization, a similar result can be obtained in the traditional regression setting. The bound on the norm of $\widehat{\alpha}$ solution of the least-squares problem may be used to derive an excess risk bound for the empirical risk minimizer in an unbounded space without truncation, at the price of a weaker dependence on $L_{\max}$, as discussed in Remark 2.

3.3 Bellman Residual Minimization and Approximation of V^π

We are now interested to relate the Bellman residual of $f_{\widehat{\alpha}}$ to the minimum Bellman residual in $\mathcal{F}$, i.e., $\inf_{f\in\mathcal{F}} \mathcal{B}(f)$, and to the approximation error (in μ-norm) of the value function V^π w.r.t. the function space $\mathcal{F}$, i.e., $\inf_{f\in\mathcal{F}} ||V^\pi - f||_\mu$. In fact, these two quantities are related since for any function $f \in \mathcal{F}$, we have $\mathcal{T}^\pi f - f = (I - \gamma P^\pi)(V^\pi - f)$. Thus, by defining

$$f_{\tilde{\alpha}} = \arg\min_{f\in\mathcal{F}} \mathcal{B}(f), \quad \text{and} \quad f_{\alpha^*} = \arg\min_{f\in\mathcal{F}} ||V^\pi - f||_\mu,$$

we have

$$||V^\pi - f_{\alpha^*}||_\mu \leqslant ||V^\pi - f_{\tilde{\alpha}}||_\mu \leqslant \frac{C^\pi(\mu)}{1-\gamma}\sqrt{\mathcal{B}(f_{\tilde{\alpha}})} \leqslant \frac{C^\pi(\mu)}{1-\gamma}\sqrt{\mathcal{B}(f_{\alpha^*})}. \tag{9.8}$$

We can now relate both the Bellman residual of $f_{\widehat{\alpha}}$, $\mathcal{B}(f_{\widehat{\alpha}})$, and its approximation error, $||V^\pi - f_{\widehat{\alpha}}||_\mu$, to the minimum possible Bellman residual in $\mathcal{F}$ and the distance between V^π and $\mathcal{F}$.

Theorem 9.2 (*Approximation error of BRM*) *For any $\delta > 0$, whenever $n \geqslant n^\pi(\nu, \delta/2)$, with probability $1-\delta$, the Bellman residual of the EBR minimizer $f_{\widehat{\alpha}}$ is bounded as*

$$\mathcal{B}(f_{\widehat{\alpha}}) \leqslant \inf_{f\in\mathcal{F}} \mathcal{B}(f) + c_4\sqrt{\frac{2d\log(2)+6\log(16/\delta)}{n}},$$

with $c_4 = (96\sqrt{2}+1)[\frac{2}{\xi}(1+\gamma)^2 L_{\max}^2+1]^2 R_{\max}^2$, and the approximation error of V^π is bounded as $||V^\pi - f_{\widehat{\alpha}}||_\mu^2 \leqslant \left(\frac{C^\pi(\mu)}{1-\gamma}\right)^2 \mathcal{B}(f_{\widehat{\alpha}})$. Moreover, since $\inf_{f\in\mathcal{F}} \mathcal{B}(f) \leqslant (1+\gamma||P^\pi||_\mu)^2 \inf_{f\in\mathcal{F}} ||V^\pi - f||_\mu^2$, we obtain an alternative bound

$$||V^\pi - f_{\widehat{\alpha}}||_\mu^2 \leqslant \left(\frac{C^\pi(\mu)}{1-\gamma}\right)^2 \left((1+\gamma||P^\pi||_\mu)^2 \inf_{f\in\mathcal{F}} ||V^\pi - f||_\mu^2 + c_4\sqrt{\frac{2d\log(2)+6\log(16/\delta)}{n}}\right).$$

Proof: From the definition of $\tilde{\alpha}$ (the minimum of $\mathcal{B}$), we have $\mathcal{L}(f_{\tilde{\alpha}}) = 2\langle R^\pi, (I-\gamma P^\pi)\varphi^\top \tilde{\alpha}\rangle_\mu$. Thus, from Lemma 9.2, we obtain

$$||\tilde{\alpha}||^2 \leqslant \frac{1}{\xi}\mathcal{L}(f_{\tilde{\alpha}}) \leqslant \frac{2}{\xi}(1+\gamma)L_{\max}R_{\max}||\tilde{\alpha}||, \text{ thus } ||\tilde{\alpha}|| \leqslant \frac{2}{\xi}(1+\gamma)L_{\max}R_{\max}.$$

Now using Chernoff Hoeffding's inequality, we have with probability $1-\delta/2$,

$$\mathcal{B}_n(f_{\tilde{\alpha}}) \leqslant \mathcal{B}(f_{\tilde{\alpha}}) + \left[\frac{2}{\xi}(1+\gamma)^2 L_{\max}^2 + 1\right]^2 R_{\max}^2\sqrt{\frac{2\log(2/\delta)}{n}}\,. \tag{9.9}$$

We may write

$$\mathcal{B}(f_{\widehat{\alpha}}) \leqslant (\mathcal{B}(f_{\widehat{\alpha}}) - \mathcal{B}_n(f_{\widehat{\alpha}})) + \mathcal{B}_n(f_{\tilde{\alpha}}) \leqslant \inf_{f\in\mathcal{F}} \mathcal{B}(f) + (\mathcal{B}(f_{\widehat{\alpha}}) - \mathcal{B}_n(f_{\widehat{\alpha}})) + (\mathcal{B}_n(f_{\tilde{\alpha}}) - \mathcal{B}(f_{\tilde{\alpha}})).$$

The claim follows by applying Theorem 9.1 (with probability $\delta/2$) and (9.9) for the second and third terms on the right hand, respectively, and a union bound so that both events hold simultaneously with probability at least $1-\delta$. The other inequalities are deduced from the definition of $\tilde{\alpha}$ and α^* and (9.8). □

This result means that whenever the space $\mathcal{F}$ is such that it contains a function with a small Bellman residual or that it can well approximate V^π, then the residual of the EBR minimizer $f_{\widehat{\alpha}}$ is small. In addition, assuming that $C^\pi(\mu)$ is small, $f_{\widehat{\alpha}}$ is also a good approximation of the value function V^π.

Input: Function space $\mathcal{F}$, state distribution μ, number of samples n, number of iterations K
Initialize: Let $V_0 \in \mathcal{B}(\mathcal{X}; V_{\max})$ be an arbitrary value function
for $k = 1, 2, \ldots, K$ **do**
(1) Let π_k be the greedy policy w.r.t. V_{k-1} (see Eq. 9.13).
(2) Build a new dataset $\mathcal{D}_k = \{(X_i^{(k)}, R_i^{(k)}, Y_i^{(k)}, Y_i^{'(k)})\}_{i=1}^n$, where $X_i^{(k)} \overset{\text{iid}}{\sim} \mu$, $R_i^{(k)} = r(X_i^{(k)}, \pi_k(X_i^{(k)}))$, and use the generative model to draw two independent samples $Y_i^{(k)}$ and $Y_i^{'(k)}$ from $P^{\pi_k}(\cdot|X_i^{(k)})$.
(3) Let $\widehat{\alpha}_k$ be the solution to the linear system $A_k\alpha = b_k$, where A_k and b_k are defined by (9.10) and (9.11).
(4) Let $V_k = f_{\widehat{\alpha}_k}$.
Return policy π_K.

Figure 9.1: The Bellman Residual Minimization Policy Iteration (BRM-PI) algorithm.

4 Bellman Residual Minimization for Policy Iteration

We now move to the full analysis of the policy iteration algorithm where at each iteration k, the policy π_k is approximated by the solution of an empirical Bellman residual minimization. The Bellman Residual Minimization Policy Iteration (BRM-PI) algorithm is described in Figure 9.1. At each iteration k, BRM-PI generates a new dataset $\mathcal{D}_k = \{(X_i^{(k)}, R_i^{(k)}, Y_i^{(k)}, Y_i^{'(k)})\}_{i=1}^n$ where $X_i^{(k)} \overset{\text{iid}}{\sim} \mu$, $R_i^{(k)} = r(X_i^{(k)}, \pi_k(X_i^{(k)}))$, and $Y_i^{(k)}$ and $Y_i^{'(k)}$ are two independent samples drawn from $P^{\pi_k}(\cdot|X_i^{(k)})$. The $d \times d$-matrix A_k and d-vector b_k are defined as

$$A_k = \frac{1}{n}(\Psi_k^\top \Psi'_k + \Psi_k'^\top \Psi_k) \tag{9.10}$$

$$b_k = \frac{1}{n}(\Psi_k + \Psi'_k)^\top R^{(k)} \tag{9.11}$$

where $(\Psi_k)_{ij} = \varphi_j(X_i^{(k)}) - \gamma\varphi_j(Y_i^{(k)})$ and $(\Psi'_k)_{ij} = \varphi_j(X_i^{(k)}) - \gamma\varphi_j(Y_i^{'(k)})$. Then $\widehat{\alpha}_k$ is defined as the solution of

$$A_k\alpha = b_k \tag{9.12}$$

(the next theorem will provide conditions under which this system has a solution), which defines the approximation $V_k = f_{\widehat{\alpha}_k}$ of the current value function V^{π_k}. Finally, the approximation V_k is used to generate the policy π_{k+1} for the next iteration $k+1$

$$\pi_{k+1}(x) = \arg\max_{a \in \mathcal{A}} \Big[r(x,a) + \gamma \int_{\mathcal{X}} P(dy|x,a) V_k(y) \Big]. \tag{9.13}$$

Note that in order to compute the expectation we can use the generative model and replace the expectation by an average over a sufficiently large number of samples. However

this is not convenient and a usual technique used to avoid computing the expectations for deriving the greedy policy is to use action-value functions Q instead of value functions V (see e.g., Watkins 1989, Lagoudakis and Parr 2003, Antos et al. 2008), or functions defined over post-decision states (Powell, 2007). We do not further develop this point here but we simply mention that all the finite-sample analysis derived in the previous section for the setting of value functions can be easily extended to action-value functions.

Now following the analysis of Munos (2003) and Antos et al. (2008), we relate the performance of the policy π_K returned by the algorithm to the optimal policy $||V^* - V^{\pi_K}||_\rho$ (where ρ is a distribution chosen by the user), in terms of the Bellman residuals of the EBR minimizers $f_{\widehat{\alpha}_k}$ at all the iterations $k < K$ of the BRM-PI algorithm. In order to do so, we make use of the concentrability coefficients, $C_{\rho,\mu}$, defined for any couple of distributions ρ and μ in Antos et al. (2008) and Munos and Szepesvári (2008) (A refined analysis can be found in Farahmand et al. (2010)).

Let us also define $n(\delta) = \sup_\pi n^\pi(\nu^\pi, \delta)$ and write $\mathcal{B}^\pi(f) = ||f - \mathcal{T}^\pi f||_\mu^2$ the Bellman residual of f under policy π. We can now state the main result which provides a performance bound for BRM-PI.

Theorem 9.3 (*Performance bound of BRM-PI*) *For any $\delta > 0$, whenever $n \geqslant n(\delta/K)$, with probability $1-\delta$, the EBR minimizer $f_{\widehat{\alpha}_k}$, where $\widehat{\alpha}_k$ is the solution of the linear system (9.12), exists for all iterations $1 \leqslant k < K$, thus the BRM-PI algorithm is well defined, and the performance V^{π_K} of the policy π_K returned by the algorithm is such that*

$$||V^* - V^{\pi_K}||_\rho^2 \leqslant \left(\frac{2\gamma}{(1-\gamma)^2}\right)^2 \left[C_{\rho,\mu} \sup_{1\leqslant k<K} \left(\inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f) + c_k \sqrt{\frac{2d\log(2) + 6\log(16K/\delta)}{n}} \right) + \gamma^K R_{\max}^2\right],$$

where $c_k = (96\sqrt{2}+1)[\frac{2}{\xi_k}(1+\gamma)^2 L_{\max}^2 + 1]^2 R_{\max}^2$, with ξ_k defined similarly as ξ in Lemma 9.2 for the policy π_k. A bound using the distances between the sequence of value functions and $\mathcal{F}$ can be obtained using the fact that $\inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f) \leqslant (1+\gamma||P^{\pi_k}||_\mu)^2 \inf_{f\in\mathcal{F}} ||V^{\pi_k} - f||_\mu^2$.

Proof: From Antos et al. (2008, Lemma 12) we have

$$||V^* - V^{\pi_K}||_\rho^2 \leqslant \left(\frac{2\gamma}{(1-\gamma)^2}\right)^2 \left(C_{\rho,\mu} \max_{0\leqslant k<K} \mathcal{B}^{\pi_k}(f_{\widehat{\alpha}_k}) + \gamma^K R_{\max}^2\right). \tag{9.14}$$

Now from Lemma 9.4, we have that at each step $k < K$, whenever $n \geqslant n(\delta/K) \geqslant n^{\pi_k}(\nu^{\pi_k}, \delta/K)$, with probability $1 - \delta/K$, the EBR minimizer $f_{\widehat{\alpha}_k}$ exists and from Theorem 9.2, the Bellman residual of $f_{\widehat{\alpha}_k}$ is bounded as

$$\mathcal{B}^{\pi_k}(f_{\widehat{\alpha}_k}) \leqslant \inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f) + c_k \sqrt{\frac{2d\log(2) + 6\log(16K/\delta)}{n}},$$

where we used a union bound that guarantees that these bounds hold for all K iterations. □

The performance bounds reported in Theorem 9.3 are composed of the sum of three terms. The first term is an approximation error term, which indicates how well the function space $\mathcal{F}$ is adapted to the problem, either in terms of containing functions with low Bellman residuals (for the sequence of policies) $\inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f)$, or in terms of well approximating the corresponding value functions $\inf_{f\in\mathcal{F}} ||V^{\pi_k} - f||_\mu$. The second term is an estimation error term, which decreases as $O(1/\sqrt{n})$, and the third term is decreasing exponentially fast with K, the number of policy iterations.

Remark: In the current description of the BRM-PI algorithm, we regenerate a new dataset $\mathcal{D}_k$ at each policy evaluation step. However, we could generate once for all n samples $(X_1, \ldots, X_n)$ and all actions $a \in \mathcal{A}$, the corresponding rewards $R_i(a) = r(X_i, a)$ and $2n$ independent next states $Y_i(a)$ and $Y_i'(a)$ sampled from $P(\cdot|X_i, a)$. Then at each iteration k, we use these samples and build the dataset $\mathcal{D}_k = \left\{\left(X_i, R_i(\pi_k(X_i)), Y_i(\pi_k(X_i)), Y_i'(\pi_k(X_i))\right)\right\}_{i=1}^n$. This sampling strategy requires generating $2n\times|\mathcal{A}|$ samples instead of $2n\times K$ for the previous method, which is advantageous when $|\mathcal{A}| \leqslant K$. In terms of performance, this version attains a similar performance as in Theorem 9.3. The main difference is that at each iteration k, the target function V^{π_k} depends on the samples because the policy π_k is greedy w.r.t. the function $f_{\alpha_{k-1}}$ learned at the previous iteration. As a result, Lemma 9.1 should be restated by taking a supremum over all the possible policies that can be generated as greedy policies of the functions in $\mathcal{F}$. The complexity of this space of policies depends on the number of actions $|\mathcal{A}|$ and the dimension d. Finally, the complexity of the joint space obtained by $\mathcal{F}$ and the space of policies would appear in the final bound which would differ from the one in Theorem 9.3 only in constant factors.

5 Conclusion and comparison with LSTD

We provided a generalization bound for Bellman residuals and used it to provide performance bounds for an approximate policy iteration algorithm in which an empirical Bellman residual minimization is used at each policy evaluation step.

Compared to the LSTD approach analyzed in Lazaric et al. (2010c) we have a poorer estimation rate of $O(1/\sqrt{n})$ instead of $O(1/n)$ and it is an open question to whether an improved rate for Bellman residuals can be obtained (see Remark 1). The assumptions are also different: in this BRM approach we assumed that we have a generative model and thus performance bounds can be obtained under any sampling distribution μ, whereas since LSTD only requires the observation of a single trajectory (following a given policy) it can only provide performance bounds under the stationary distribution of that policy. However in a policy iteration scheme it is not enough to accurately approximate the current policy under the stationary distribution since the greedy policy w.r.t. that approximation can be arbitrarily poor. Thus the performance of BRM are better controlled than that of LSTD, which is reflected in the fact that the concentrability coefficients $C(\rho, \mu)$ (used in Theorem 9.3)

can be controlled in the BRM approach (such as by choosing a uniform distribution μ) but not in LSTD unless we make additional (usually strong) assumptions on the stationary distributions (such as being lower-bounded by a uniform distribution, like in (Munos, 2003)).

6 Technical details

6.1 Proof of Lemma 9.3

Step 1: Introduce the empirical process. Let $\mathcal{J}(C)$ be the class of functions induced by $\mathcal{L}_n$ from $\mathcal{F}(C)$ defined as

$$\mathcal{J}(C) = \{j_\alpha : (x,y,z) \mapsto (f_\alpha(x) - \gamma f_\alpha(y))(f_\alpha(x) - \gamma f_\alpha(z)); ||\alpha||_2 \leqslant C\}.$$

Note that this is the product of two linear spaces of dimension d. Furthermore, we can now rewrite $\mathcal{L}_n(f_\alpha) = P_n j_\alpha$ and $\mathcal{L}(f_\alpha) = P j_\alpha$, where P_n is the empirical measure w.r.t. X_i, Y_i, Y_i' and P is the measure according to which the samples are distributed. As a result both $\mathcal{L}_n(f_\alpha)$ and $\mathcal{L}(f_\alpha)$ are linear w.r.t. j_α. Note also that for any $(x,y,z) \in \mathcal{X}^3$, $|j_\alpha(x,y,z)| \leqslant ||\alpha||_2^2(1+\gamma)^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||_2^2 = C^2(1+\gamma)^2 L_{\max}^2$, using Cauchy-Schwartz's inequality.

Step 2: Bound the covering number. We want to bound the ε-covering number of the class of functions $\mathcal{J}(C)$ in norm $||.||_\infty$. Since each function j_α can be written as $j_\alpha(x,y,z) = g_\alpha(x,y)g_\alpha(x,z)$, where $g_\alpha(x,y) = \sum_{i=1}^d \alpha_i(\varphi_i(x) - \gamma\varphi_i(y))$, we can relate the covering number of $\mathcal{J}(C)$ to the covering number of the space of functions g_α. Indeed, let us consider an ε-cover G_0 for the space of functions g_α such that $||\alpha||_2 \leqslant C$. Thus for a given α there exists $g_{\alpha_0} \in G_0$ such that $||g_\alpha - g_{\alpha_0}|| \leqslant \varepsilon$. Now, we can build a cover for $\mathcal{J}(C)$. We have

$$\begin{aligned}
|j_\alpha(x,y,z) - j_{\alpha_0}(x,y,z)| &\leqslant |g_\alpha(x,y)g_\alpha(x,z) - g_{\alpha_0}(x,y)g_\alpha(x,z)| \\
&\quad + |g_{\alpha_0}(x,y)g_\alpha(x,z) - g_{\alpha_0}(x,y)g_{\alpha_0}(x,z)| \\
&\leqslant ||g_\alpha||_\infty ||g_\alpha - g_{\alpha_0}||_\infty + ||g_{\alpha_0}||_\infty ||g_\alpha - g_{\alpha_0}||_\infty \\
&\leqslant 2C(1+\gamma)L_{\max}\varepsilon,
\end{aligned}$$

which enables us to deduce that

$$\begin{aligned}
\mathcal{N}(\varepsilon, \mathcal{J}(C), ||.||_\infty) &\leqslant \mathcal{N}\Big(\frac{\varepsilon}{2C(1+\gamma)L_{\max}}, \{g_\alpha; ||\alpha||_2 \leqslant C\}, ||.||_\infty\Big) \\
&\leqslant \mathcal{N}\Big(\frac{\varepsilon}{2C(1+\gamma)L_{\max}}, \{g_\alpha; ||g||_n \leqslant C(1+\gamma)L_{\max}\}, ||.||_n\Big) \\
&\leqslant \Big(\frac{6C^2(1+\gamma)^2 L_{\max}^2}{\varepsilon}\Big)^d
\end{aligned}$$

where we used the fact that $||g||_n \leqslant ||g||_\infty$ and $||g||_n \leqslant ||\alpha||_2(1+\gamma)L_{\max}$.

Step 3: Use chaining technique. Let us consider ε_l-covers $\mathcal{J}_l$ of $\mathcal{J}(C)$, for $l = 0, \dots, \infty$, with $\mathcal{J}_0 = j_{\alpha_0}$. We moreover assume that $\mathcal{J}_{l+1}$ is a refinement of $\mathcal{J}_l$ and that $\varepsilon_{l+1} \leqslant \varepsilon_l$. Then for a given $j \in \mathcal{J}(C)$, we define $j_l = \Pi(j, \mathcal{J}_l)$ the projection of j into $\mathcal{J}_l$, for the norm $||j||_\infty$. Thus, $j = (j - j_L) + \sum_{l=1}^{L}(j_l - j_{l-1}) + j_0$. Since $0 \in \mathcal{J}(C)$, we consider $j_{\alpha_0} = 0$. Note that by definition, we need $||j||_\infty \leqslant \varepsilon_0$. Thus we define $\varepsilon_0 = C^2(1+\gamma)^2 L_{\max}^2$.

Moreover, we have for any $j \in \mathcal{J}(C)$,

$$|(P - P_n)(j)| \leqslant |(P - P_n)(j - j_L)| + \sum_{l=1}^{L} |(P - P_n)(j_l - j_{l-1})| \leqslant 2\varepsilon_L + \sum_{l=1}^{L} |(P - P_n)(j_l - j_{l-1})|$$

We introduce for convenience the following notation: $\rho(t) = \mathbb{P}(\exists f \in \mathcal{F}(C), \quad |\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| > t)$. Thus if we now introduce η and $(\eta_l)_{l \leqslant L}$ such that $\sum_{l=1}^{L} \eta_l \leqslant \eta$, then for L large enough such that $2\varepsilon_L \leqslant t_2$, we have:

$$\begin{aligned}
\rho(\eta t_1 + t_2) &\leqslant \mathbb{P}(\exists f \in \mathcal{F}(C), \quad 2\varepsilon_L + \sum_{l=1}^{L} |(P - P_n)(j_l - j_{l-1})| > \sum_{l=1}^{L} \eta_l t + t_2) \\
&\leqslant \sum_{l=1}^{L} \mathbb{P}(\exists j \in \mathcal{J}(C), \quad |(P - P_n)(j_{\alpha_l} - j_{\alpha_{l-1}})| > \eta_l t_1) \\
&\leqslant \sum_{l=1}^{L} N_l N_{l-1} \sup_{j \in \mathcal{J}(C)} \mathbb{P}(|(P - P_n)(j_{\alpha_l} - j_{\alpha_{l-1}})| > \eta_l t_1) \\
&\leqslant \sum_{l=1}^{L} 2N_l^2 \exp(-\frac{n t_1^2 \eta_l^2}{2(4\varepsilon_l)^2})
\end{aligned}$$

where $N_l = \mathcal{N}(\varepsilon_l, \mathcal{J}(C), ||.||_\infty)$, and where the last inequality comes from the fact that

$$|j_{\alpha_l}(X_i, Y_i, Y_i') - j_{\alpha_{l-1}}(X_i, Y_i, Y_i') - P j_{\alpha_l} + P j_{\alpha_{l-1}}| \leqslant 2||j_{\alpha_l} - j_{\alpha_{l-1}}||_\infty \leqslant 4||j_{\alpha_l} - j_\alpha||_\infty \leqslant 4\varepsilon_l.$$

Step 4: Define the free parameters. Thus, if we define, for all $l \geqslant 1$, $\eta_l \stackrel{\text{def}}{=} \frac{8\varepsilon_l}{t_1}\sqrt{\frac{2\log(N_l)}{n}}$, then we deduce the following inequality: $\rho(\eta t_1 + t_2) \leqslant 2\sum_{l=1}^{L} N_l^{-2}$.

Now, since $N_l \leqslant (\frac{6C^2(1+\gamma)^2 L_{\max}^2}{\varepsilon_l})^d$, let $\varepsilon_l = 6C^2(1+\gamma)^2 L_{\max}^2 2^{-l}(\delta/2)^{1/2d}(2^{2d}-1)^{1/2d}$ for

$l \geqslant 1$. Thus we deduce that $\sum_{l=1}^{L} N_l^{-2} \leqslant \delta/2$. We finally get:

$$\begin{aligned}
\eta t_1 + t_2 &= \sum_{l=1}^{L} 8\varepsilon_l \sqrt{\frac{2\log(N_l)}{n}} + 2\varepsilon_L \\
&\leqslant 48C^2(1+\gamma)^2 L_{\max}^2 (\delta/2)^{1/2d}(2^{2d}-1)^{1/2d} \sum_{l=1}^{L} 2^{-l}\sqrt{\frac{2\log(N_l)}{n}} + 2\varepsilon_L \\
&\leqslant \frac{96C^2(1+\gamma)^2 L_{\max}^2}{\sqrt{n}} \sum_{l=1}^{L} 2^{-l}\sqrt{2dl\log(2) + \log(2/\delta) - \log(2^{2d}-1)} + 2\varepsilon_L \\
&\leqslant \frac{96C^2(1+\gamma)^2 L_{\max}^2}{\sqrt{n}} \sum_{l=1}^{L} 2^{-l}\sqrt{2d(l-1)\log(2) + \log(4/\delta)} + 2\varepsilon_L
\end{aligned}$$

Thus, when $L \to \infty$, we get:

$$\eta t_1 + t_2 \leqslant \frac{96C^2(1+\gamma)^2 L_{\max}^2}{\sqrt{n}} \sum_{l=1}^{\infty} 2^{-l}\sqrt{2d(l-1)\log(2) + \log(4/\delta)}$$

We deduce that with probability higher than $1-\delta$, the following holds true:

$$\sup_{f \in \mathcal{F}(C)} |\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| \leqslant 96C^2 L_{\max}^2 \left(\sqrt{\frac{2d\log(2)}{n}} + \sqrt{\frac{\log(4/\delta)}{n}}\right)$$

Then we use the fact that $\mathcal{L}(f_\alpha) = \mathcal{L}(f_{\frac{\alpha}{||\alpha||}})||\alpha||^2$ and similarly $\mathcal{L}_n(f_\alpha) = \mathcal{L}_n(f_{\frac{\alpha}{||\alpha||}})||\alpha||^2$ to deduce that with the same probability, for all α,

$$|\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| \leqslant ||\alpha||^2 \left(\sup_{f \in \mathcal{F}(1)} |\mathcal{L}(f) - \mathcal{L}_n(f)|\right)$$

The final results follows by aesthetics simplifications.

6.2 Proof of Lemma 9.1

Step 1: Introduce the empirical process. The proof for B_n follows the same lines as for L_n using the following class of functions, induced by $\mathcal{B}_n$ from $\mathcal{F}(C)$ and defined as:

$$\mathcal{J}(C) = \{j_\alpha : (x,y,z) \mapsto (f_\alpha(x) - \gamma f_\alpha(y) + r(x))(f_\alpha(x) - \gamma f_\alpha(z) + r(x)); ||\alpha||_2 \leqslant C\}.$$

Then we have $\mathcal{B}_n(f_\alpha) = P_n j_\alpha$ and $\mathcal{B}(f_\alpha) = P j_\alpha$. Now, we have $|j_\alpha(X_i, Y_i, Y_i')| \leqslant (||\alpha||_2(1+\gamma)\sup_x ||\varphi(x)||_2 + R_{\max})^2 = [C(1+\gamma)L_{\max} + R_{\max}]^2$. Note that the function 0 does not a priori belongs to $\mathcal{J}(C)$, thus we have an additional term to control corresponding to the decomposition of $j = (j - j_L) + \sum_{l=1}^{L}(j_l - j_{l-1}) + j_0$ for some nonzero $j_0 \in J(C)$.

Step 2: Bound the covering number. With this new definition of $\mathcal{J}(C)$, we have:

$$\mathcal{N}(\varepsilon, \mathcal{J}(C), ||.||_\infty) \leqslant \Big(\frac{6(C(1+\gamma)L_{\max} + R_{\max})^2}{\varepsilon_l}\Big)^d$$

Step 3: Use chaining technique. Then using chaining technique, we get the corresponding upper bound:

$$\begin{aligned}\rho(\eta t_1 + t_2 + t_3) &= \mathbb{P}(\exists f \in \mathcal{F}(C)|\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| > \eta t_1 + t_2 + t_3) \\ &\leqslant 2\sum_{l=1}^{L} N_l^{-2} + 2\exp(-\frac{nt_3^2}{2[C(1+\gamma)L_{\max} + R]^4})\end{aligned}$$

where the last term comes from the bound on $\mathbb{P}(|(P - P_n)(j_0) \geqslant t_3)$.

Step 4: Define the free parameters. We define $\varepsilon_l \stackrel{\text{def}}{=} 9(C(1+\gamma)L_{\max}+R)^2 2^{-l}(\delta/4)^{1/2d}(2^{2d}-1)^{1/2d}$ for $l \geqslant 1$, set $t_3 = [C(1+\gamma)L_{\max}+R]^2\sqrt{\frac{2\log(4/\delta)}{n}}$ and derive that with probability higher than $1-\delta$,

$$\begin{aligned}\sup_{f\in\mathcal{F}(C)} |\mathcal{B}(f_\alpha) - \mathcal{B}_n(f_\alpha)| &\leqslant 96[C(1+\gamma)L_{\max} + R]^2(\sqrt{\frac{2d\log(2)}{n}} + \sqrt{\frac{\log(8/\delta)}{n}}) \\ &\quad + [C(1+\gamma)L_{\max} + R]^2\sqrt{\frac{2\log(4/\delta)}{n}}.\end{aligned}$$

The final result follows after some aesthetics simplifications.

Chapter 10

Least-squares TD with Random Projections.

We consider the problem of reinforcement learning in high-dimensional spaces when the number of features is bigger than the number of samples. In particular, we study the least-squares temporal difference (LSTD) learning algorithm when a space of low dimension is generated with a random projection from a high-dimensional space. We provide a thorough theoretical analysis of the LSTD with random projections and derive performance bounds for the resulting algorithm. We also show how the error of LSTD with random projections is propagated through the iterations of a policy iteration algorithm and provide a performance bound for the resulting least-squares policy iteration (LSPI) algorithm.

The work presented in this chapter is a joint work with *Mohammad Ghavamzadeh* and *Alessandro Lazaric* and has been published in the proceedings of the 24*th conference on advances in Neural Information Processing Systems (NIPS 2010)*, see Ghavamzadeh et al. (2010a).

Contents

1 Introduction

Least-squares temporal difference (LSTD) learning Bradtke and Barto (1996), Boyan (1999) is a widely used reinforcement learning (RL) algorithm for learning the value function V^π of a given policy π. LSTD has been successfully applied to a number of problems especially after the development of the least-squares policy iteration (LSPI) algorithm Lagoudakis and Parr (2003), which extends LSTD to control problems by using it in the policy evaluation step of policy iteration. More precisely, LSTD computes the fixed point of the operator $\Pi\mathcal{T}^\pi$, where $\mathcal{T}^\pi$ is the Bellman operator of policy π and Π is the projection operator onto a linear function space. The choice of the linear function space has a major impact on the accuracy of the value function estimated by LSTD, and thus, on the quality of the policy learned by LSPI. The problem of finding the right space, or in other words the problems of feature selection and discovery, is an important challenge in many areas of machine learning including RL, or more specifically, linear value function approximation in RL.

To address this issue in RL, many researchers have focused on feature extraction and learning. Mahadevan Mahadevan (2005) proposed a constructive method for generating features based on the eigenfunctions of the Laplace-Beltrami operator of the graph built from observed system trajectories. Menache et al. Menache et al. (2005) presented a method that starts with a set of features and then tunes both features and the weights using either gradient descent or the cross-entropy method. Keller et al. Keller et al. (2006) proposed an algorithm in which the state space is repeatedly projected onto a lower dimensional space based on the Bellman error and then states are aggregated in this space to define new features. Finally, Parr et al. Parr et al. (2007) presented a method that iteratively adds features to a linear approximation architecture such that each new feature is derived from the Bellman error of the existing set of features.

A more recent approach to feature selection and discovery in value function approximation in RL is to solve *RL in high-dimensional feature spaces*. The basic idea here is to use a large number of features and then exploit the regularities in the problem to solve it efficiently in this high-dimensional space. Theoretically speaking, increasing the size of the function space can reduce the approximation error (the distance between the target function and the space) at the cost of a growth in the estimation error. In practice, in the typical high-dimensional learning scenario when the number of features is larger than the number of samples, this often leads to the overfitting problem and poor prediction performance. To overcome this problem, several approaches have been proposed including regularization. Both ℓ_1 and ℓ_2 regularizations have been studied in value function approximation in RL. Farahmand et al. presented several ℓ_2-regularized RL algorithms by adding ℓ_2-regularization to LSTD and modified Bellman residual minimization Farahmand et al. (2008) as well as fitted value iteration Farahmand et al. (2009), and proved finite-sample performance bounds for their algorithms. There have also been algorithmic work on adding ℓ_1-penalties to the TD Loth et al. (2007), LSTD Kolter and Ng (2009), and linear programming Petrik et al. (2010) algorithms.

In this paper, we follow a different approach based on random projections Vempala (2004).

In particular, we study the performance of *LSTD with random projections* (LSTD-RP). Given a high-dimensional linear space $\mathcal{F}$, LSTD-RP learns the value function of a given policy from a small (relative to the dimension of $\mathcal{F}$) number of samples in a space $\mathcal{G}$ of lower dimension obtained by linear random projection of the features of $\mathcal{F}$. We prove that solving the problem in the low dimensional random space instead of the original high-dimensional space reduces the estimation error at the price of a "controlled" increase in the approximation error of the original space $\mathcal{F}$. We present the LSTD-RP algorithm and discuss its computational complexity in Section 3. In Section 4, we provide the finite-sample analysis of the algorithm. Finally in Section 5, we show how the error of LSTD-RP is propagated through the iterations of LSPI.

2 Preliminaries

For a measurable space with domain $\mathcal{X}$, we let $\mathcal{S}(\mathcal{X})$ and $\mathcal{B}(\mathcal{X}; L)$ denote the set of probability measures over $\mathcal{X}$ and the space of measurable functions with domain $\mathcal{X}$ and bounded in absolute value by $0 < L < \infty$, respectively. For a measure $\mu \in \mathcal{S}(\mathcal{X})$ and a measurable function $f : \mathcal{X} \to \mathbb{R}$, we define the $\ell_2(\mu)$-norm of f as $||f||_\mu^2 = \int f(x)^2 \mu(dx)$, the supremum norm of f as $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$, and for a set of n states $X_1, \ldots, X_n \in \mathcal{X}$ the empirical norm of f as $||f||_n^2 = \frac{1}{n} \sum_{t=1}^n f(X_t)^2$. Moreover, for a vector $u \in \mathbb{R}^n$ we write its ℓ_2-norm as $||u||_2^2 = \sum_{i=1}^n u_i^2$.

We consider the standard RL framework Sutton and Barto (1998) in which a learning agent interacts with a stochastic environment and this interaction is modeled as a discrete-time discounted Markov decision process (MDP). A discount MDP is a tuple $\mathcal{M} = \langle \mathcal{X}, \mathcal{A}, r, P, \gamma \rangle$ where the state space $\mathcal{X}$ is a bounded closed subset of a Euclidean space, $\mathcal{A}$ is a finite ($|\mathcal{A}| < \infty$) action space, the reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is uniformly bounded by $R_{\max}$, the transition kernel P is such that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $P(\cdot|x, a)$ is a distribution over $\mathcal{X}$, and $\gamma \in (0, 1)$ is a discount factor. A deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$ is a mapping from states to actions. Under a policy π, the MDP $\mathcal{M}$ is reduced to a Markov chain $\mathcal{M}^\pi = \langle \mathcal{X}, R^\pi, P^\pi, \gamma \rangle$ with reward $R^\pi(x) = r\big(x, \pi(x)\big)$, transition kernel $P^\pi(\cdot|x) = P\big(\cdot\,|x, \pi(x)\big)$, and stationary distribution ρ^π (if it admits one). The value function of a policy π, V^π, is the unique fixed-point of the Bellman operator $\mathcal{T}^\pi : \mathcal{B}(\mathcal{X}; V_{\max} = \frac{R_{\max}}{1-\gamma}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by $(\mathcal{T}^\pi V)(x) = R^\pi(x) + \gamma \int_{\mathcal{X}} P^\pi(dy|x) V(y)$. We also define the optimal value function V^* as the unique fixed-point of the optimal Bellman operator $\mathcal{T}^* : \mathcal{B}(\mathcal{X}; V_{\max}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by $(\mathcal{T}^* V)(x) = \max_{a \in \mathcal{A}} \big[r(x, a) + \gamma \int_{\mathcal{X}} P(dy|x, a) V(y)\big]$. Finally, we denote by T the truncation operator at threshold $V_{\max}$, i.e., if $|f(x)| > V_{\max}$ then $T(f)(x) = \operatorname{sgn}\big(f(x)\big) V_{\max}$.

To approximate a value function $V \in \mathcal{B}(\mathcal{X}; V_{\max})$, we first define a linear function space $\mathcal{F}$ spanned by the basis functions $\varphi_j \in \mathcal{B}(\mathcal{X}; L)$, $j = 1, \ldots, D$, i.e., $\mathcal{F} = \{f_\alpha \mid f_\alpha(\cdot) = \varphi(\cdot)^\top \alpha,\ \alpha \in \mathbb{R}^D\}$, where $\varphi(\cdot) = \big(\varphi_1(\cdot), \ldots, \varphi_D(\cdot)\big)^\top$ is the feature vector. We define the orthogonal projection of V onto the space $\mathcal{F}$ w.r.t. norm μ as $\Pi_{\mathcal{F}} V = \arg\min_{f \in \mathcal{F}} ||V - f||_\mu$. From $\mathcal{F}$ we can generate a d-dimensional ($d < D$) random space $\mathcal{G} = \{g_\beta \mid g_\beta(\cdot) = \Psi(\cdot)^\top \beta,\ \beta \in \mathbb{R}^d\}$, where the feature vector $\Psi(\cdot) = \big(\psi_1(\cdot), \ldots, \psi_d(\cdot)\big)^\top$ is defined as $\Psi(\cdot) =$

$A\varphi(\cdot)$ with $A \in \mathbb{R}^{d\times D}$ be a random matrix whose elements are drawn i.i.d. from a suitable distribution, e.g., Gaussian $\mathcal{N}(0, 1/d)$. Similar to the space $\mathcal{F}$, we define the orthogonal projection of V onto the space $\mathcal{G}$ w.r.t. norm μ as $\Pi_{\mathcal{G}} V = \arg\min_{g\in\mathcal{G}} ||V - g||_\mu$. Finally, for any function $f_\alpha \in \mathcal{F}$, we define $m(f_\alpha) = ||\alpha||_2 \sup_{x\in\mathcal{X}} ||\varphi(x)||_2$.

3 LSTD with Random Projections

The objective of LSTD with random projections (LSTD-RP) is to learn the value function of a given policy from a small (relative to the dimension of the original space) number of samples in a low-dimensional linear space defined by a random projection of the high-dimensional space. We show that solving the problem in the low dimensional space instead of the original high-dimensional space reduces the estimation error at the price of a "controlled" increase in the approximation error. In this section, we introduce the notations and the resulting algorithm, and discuss its computational complexity. In Section 4, we provide the finite-sample analysis of the algorithm.

We use the linear spaces $\mathcal{F}$ and $\mathcal{G}$ with dimensions D and d ($d < D$) as defined in Section 2. Since in the following the policy is fixed, we drop the dependency of R^π, P^π, V^π, and $\mathcal{T}^\pi$ on π and simply use R, P, V, and $\mathcal{T}$. Let $\{X_t\}_{t=1}^n$ be a sample path (or trajectory) of size n generated by the Markov chain $\mathcal{M}^\pi$, and let $v \in \mathbb{R}^n$ and $r \in \mathbb{R}^n$, defined as $v_t = V(X_t)$ and $r_t = R(X_t)$, be the value and reward vectors of this trajectory. Also, let $\Psi = [\psi(X_1)^\top; \dots; \psi(X_n)^\top]$ be the feature matrix defined at these n states and $\mathcal{G}_n = \{\Psi\beta \mid \beta \in \mathbb{R}^d\} \subset \mathbb{R}^n$ be the corresponding vector space. We denote by $\widehat{\Pi}_{\mathcal{G}} : \mathbb{R}^n \to \mathcal{G}_n$ the orthogonal projection onto $\mathcal{G}_n$, defined by $\widehat{\Pi}_{\mathcal{G}} y = \arg\min_{z\in\mathcal{G}_n} ||y - z||_n$, where $||y||_n^2 = \frac{1}{n}\sum_{t=1}^n y_t^2$. Similarly, we can define the orthogonal projection onto $\mathcal{F}_n = \{\Phi\alpha \mid \alpha \in \mathbb{R}^D\}$ as $\widehat{\Pi}_{\mathcal{F}} y = \arg\min_{z\in\mathcal{F}_n} ||y - z||_n$, where $\Phi = [\varphi(X_1)^\top; \dots; \varphi(X_n)^\top]$ is the feature matrix defined at $\{X_t\}_{t=1}^n$. Note that for any $y \in \mathbb{R}^n$, the orthogonal projections $\widehat{\Pi}_{\mathcal{G}} y$ and $\widehat{\Pi}_{\mathcal{F}} y$ exist and are unique.

We consider the pathwise-LSTD algorithm introduced in Lazaric et al. (2010a). Pathwise-LSTD takes a single trajectory $\{X_t\}_{t=1}^n$ of size n generated by the Markov chain as input and returns the fixed point of the empirical operator $\widehat{\Pi}_{\mathcal{G}}\widehat{\mathcal{T}}$, where $\widehat{\mathcal{T}}$ is the pathwise Bellman operator defined as $\widehat{\mathcal{T}} y = r + \gamma \widehat{P} y$. The operator $\widehat{P} : \mathbb{R}^n \to \mathbb{R}^n$ is defined as $(\widehat{P}y)_t = y_{t+1}$ for $1 \leqslant t < n$ and $(\widehat{P}y)_n = 0$. As shown in Lazaric et al. (2010a), $\widehat{\mathcal{T}}$ is a γ-contraction in ℓ_2-norm, thus together with the non-expansive property of $\widehat{\Pi}_{\mathcal{G}}$, it guarantees the existence and uniqueness of the pathwise-LSTD fixed point $\widehat{v} \in \mathbb{R}^n$, $\widehat{v} = \widehat{\Pi}_{\mathcal{G}}\widehat{\mathcal{T}}\widehat{v}$. Note that the uniqueness of $\widehat{v}$ does not imply the uniqueness of the parameter $\widehat{\beta}$ such that $\widehat{v} = \Psi\widehat{\beta}$.

Figure 10.1 contains the pseudo-code and the computational cost of the LSTD-RP algorithm. The total computational cost of LSTD-RP is $O(d^3 + ndD)$, while the computational cost of LSTD in the high-dimensional space $\mathcal{F}$ is $O(D^3 + nD^2)$. As we will see, the analysis of Section 4 suggests that the value of d should be set to $O(\sqrt{n})$. In this case the numerical complexity of LSTD-RP is $O(n^{3/2}D)$, which is better than $O(D^3)$, the cost of LSTD in $\mathcal{F}$ when $n < D$ (the case considered in this paper). Note that the cost of making a prediction

LSTD-RP $\left(D, d, \{X_t\}_{t=1}^n, \{R(X_t)\}_{t=1}^n, \varphi, \gamma\right)$ **Cost**
Compute

- the reward vector $r_{n\times 1}$; $r_t = R(X_t)$ $O(n)$
- the high-dimensional feature matrix $\Phi_{n\times D} = [\varphi(X_1)^\top; \ldots; \varphi(X_n)^\top]$ $O(nD)$
- the projection matrix $A_{d\times D}$ whose elements are i.i.d. samples from $\mathcal{N}(0, 1/d)$ $O(dD)$
- the low-dim feature matrix $\Psi_{n\times d} = [\psi(X_1)^\top; \ldots; \psi(X_n)^\top]$; $\psi(\cdot) = A\varphi(\cdot)$ $O(ndD)$
- the matrix $\widehat{P}\Psi = \Psi'_{n\times d} = [\psi(X_2)^\top; \ldots; \psi(X_n)^\top; \mathbf{0}^\top]$ $O(nd)$
- $\tilde{A}_{d\times d} = \Psi^\top(\Psi - \gamma\Psi')$, $\tilde{b}_{d\times 1} = \Psi^\top r$ $O(nd + nd^2) + O(nd)$

return either $\widehat{\beta} = \tilde{A}^{-1}\tilde{b}$ or $\widehat{\beta} = \tilde{A}^{+}\tilde{b}$
($\tilde{A}^+$ is the Moore-Penrose pseudo-inverse of $\tilde{A}$) $O(d^2 + d^3)$

Figure 10.1: The pseudo-code of the LSTD with random projections (LSTD-RP) algorithm.

is D in LSTD in $\mathcal{F}$ and dD in LSTD-RP.

4 Finite-Sample Analysis of LSTD with Random Projections

In this section, we report the main theoretical results of the paper. In particular, we derive a performance bound for LSTD-RP in the Markov design setting, i.e., when the LSTD-RP solution is compared to the true value function only at the states belonging to the trajectory used by the algorithm (see Section 4 in Lazaric et al. (2010a) for a more detailed discussion). We then derive a condition on the number of samples to guarantee the uniqueness of the LSTD-RP solution. Finally, from the Markov design bound we obtain generalization bounds when the Markov chain has a stationary distribution.

4.1 Markov Design Bound

Theorem 10.1 (*Performance bound of LSTD-RP with Markov design*) *Let $\mathcal{F}$ and $\mathcal{G}$ be linear spaces with dimensions D and d ($d < D$) as defined in Section 2. Let $\{X_t\}_{t=1}^n$ be a sample path generated by the Markov chain $\mathcal{M}^\pi$, and $v, \widehat{v} \in \mathbb{R}^n$ be the vectors whose components are the value function and the LSTD-RP solution at $\{X_t\}_{t=1}^n$. Then for any $\delta > 0$, whenever $d \geqslant 15 \log(8n/\delta)$, with probability $1 - \delta$ (the randomness is w.r.t. both the*

random sample path and the random projection), $\widehat{v}$ satisfies

$$||v-\widehat{v}||_n \leqslant \frac{1}{\sqrt{1-\gamma^2}}\left[||v-\widehat{\Pi}_{\mathcal{F}}v||_n + \sqrt{\frac{8\log(8n/\delta)}{d}}m(\widehat{\Pi}_{\mathcal{F}}v)\right] + \frac{\gamma V_{\max}L}{1-\gamma}\sqrt{\frac{d}{\nu_n}}\left(\sqrt{\frac{8\log(4d/\delta)}{n}}+\frac{1}{n}\right), \tag{10.1}$$

where the random variable ν_n is the smallest strictly positive eigenvalue of the sample-based Gram matrix $\frac{1}{n}\Psi^\top\Psi$. Note that $m(\widehat{\Pi}_{\mathcal{F}}v) = m(f_\alpha)$ with f_α be any function in $\mathcal{F}$ such that $f_\alpha(X_t) = (\widehat{\Pi}_{\mathcal{F}}v)_t$ for $1 \leqslant t \leqslant n$.

Before stating the proof of Theorem 10.1, we need to prove the following lemma.

Lemma 10.1 *Let $\mathcal{F}$ and $\mathcal{G}$ be linear spaces with dimensions D and d ($d < D$) as defined in Section 2. Let $\{X_i\}_{i=1}^n$ be n states and $f_\alpha \in \mathcal{F}$. Then for any $\delta > 0$, whenever $d \geqslant 15\log(4n/\delta)$, with probability $1-\delta$ (the randomness is w.r.t. the random projection), we have*

$$\inf_{g\in\mathcal{G}} ||f_\alpha - g||_n^2 \leqslant \frac{8\log(4n/\delta)}{d}m(f_\alpha)^2. \tag{10.2}$$

Proof: The proof relies on the application of a variant of Johnson-Lindenstrauss (JL) lemma which states that the inner-products are approximately preserved by the application of the random matrix A (see e.g., Proposition 1 in Maillard and Munos (2009)). For any $\delta > 0$, we set $\varepsilon^2 = \frac{8}{d}\log(4n/\delta)$. Thus for $d \geqslant 15\log(4n/\delta)$, we have $\varepsilon \leqslant 3/4$ and as a result $\varepsilon^2/4 - \varepsilon^3/6 \geqslant \varepsilon^2/8$ and $d \geqslant \frac{\log(4n/\delta)}{\varepsilon^2/4-\varepsilon^3/6}$. Thus, from Proposition 1 in Maillard and Munos (2009), for all $1 \leqslant i \leqslant n$, we have $|\varphi(X_i)\cdot\alpha - A\varphi(X_i)\cdot A\alpha| \leqslant \varepsilon||\alpha||_2||\varphi(X_i)||_2 \leqslant \varepsilon\, m(f_\alpha)$ with high probability. From this result, we deduce that with probability $1-\delta$

$$\inf_{g\in\mathcal{G}} ||f_\alpha - g||_n^2 \leqslant ||f_\alpha - g_{A\alpha}||_n^2 = \frac{1}{n}\sum_{i=1}^n |\varphi(X_i)\cdot\alpha - A\varphi(X_i)\cdot A\alpha|^2 \leqslant \frac{8\log(4n/\delta)}{d}m(f_\alpha)^2.$$

Proof: [Proof of Theorem 10.1] For any fixed space $\mathcal{G}$, the performance of the LSTD-RP solution can be bounded according to Theorem 1 in Lazaric et al. (2010b) as

$$||v-\widehat{v}||_n \leqslant \frac{1}{\sqrt{1-\gamma^2}}||v-\widehat{\Pi}_{\mathcal{G}}v||_n + \frac{\gamma V_{\max}L}{1-\gamma}\sqrt{\frac{d}{\nu_n}}\Big(\sqrt{\frac{8\log(2d/\delta')}{n}}+\frac{1}{n}\Big), \tag{10.3}$$

with probability $1-\delta'$ (w.r.t. the random sample path). From the triangle inequality, we have

$$||v-\widehat{\Pi}_{\mathcal{G}}v||_n \leqslant ||v-\widehat{\Pi}_{\mathcal{F}}v||_n + ||\widehat{\Pi}_{\mathcal{F}}v-\widehat{\Pi}_{\mathcal{G}}v||_n = ||v-\widehat{\Pi}_{\mathcal{F}}v||_n + ||\widehat{\Pi}_{\mathcal{F}}v-\widehat{\Pi}_{\mathcal{G}}(\widehat{\Pi}_{\mathcal{F}}v)||_n. \tag{10.4}$$

The equality in Eq. 10.4 comes from the fact that for any vector $g \in \mathcal{G}$, we can write $||v-g||_n^2 = ||v-\widehat{\Pi}_{\mathcal{F}}v||_n^2 + ||\widehat{\Pi}_{\mathcal{F}}v-g||_n^2$. Since $||v-\widehat{\Pi}_{\mathcal{F}}v||_n$ is independent of g, we have $\arg\inf_{g\in\mathcal{G}}||v-g||_n^2 = \arg\inf_{g\in\mathcal{G}}||\widehat{\Pi}_{\mathcal{F}}v-g||_n^2$, and thus, $\widehat{\Pi}_{\mathcal{G}}v = \widehat{\Pi}_{\mathcal{G}}(\widehat{\Pi}_{\mathcal{F}}v)$. From Lemma 10.1, if $d \geqslant 15\log(4n/\delta'')$, with probability $1-\delta''$ (w.r.t. the choice of A), we have

$$||\widehat{\Pi}_{\mathcal{F}}v-\widehat{\Pi}_{\mathcal{G}}(\widehat{\Pi}_{\mathcal{F}}v)||_n \leqslant \sqrt{\frac{8\log(4n/\delta'')}{d}}m(\widehat{\Pi}_{\mathcal{F}}v). \tag{10.5}$$

□

We conclude from a union bound argument that Eqs. 10.3 and 10.5 hold simultaneously with probability at least $1 - \delta' - \delta''$. The claim follows by combining Eqs. 10.3–10.5, and setting $\delta' = \delta'' = \delta/2$. □

Remark 1. Using Theorem 10.1, we can compare the performance of LSTD-RP with the performance of LSTD directly applied in the high-dimensional space $\mathcal{F}$. Let $\bar{v}$ be the LSTD solution in $\mathcal{F}$, then up to constants, logarithmic, and dominated factors, with high probability, $\bar{v}$ satisfies

$$||v - \bar{v}||_n \leqslant \frac{1}{\sqrt{1-\gamma^2}}||v - \widehat{\Pi}_{\mathcal{F}} v||_n + \frac{1}{1-\gamma} O(\sqrt{D/n}). \tag{10.6}$$

By comparing Eqs. 10.1 and 10.6, we notice that **1)** the estimation error of $\widehat{v}$ is of order $O(\sqrt{d/n})$, and thus, is smaller than the estimation error of $\bar{v}$, which is of order $O(\sqrt{D/n})$, and **2)** the approximation error of $\widehat{v}$ is the approximation error of $\bar{v}$, $||v - \widehat{\Pi}_{\mathcal{F}} v||_n$, plus an additional term that depends on $m(\widehat{\Pi}_{\mathcal{F}} v)$ and decreases with d, the dimensionality of $\mathcal{G}$, with the rate $O(\sqrt{1/d})$. Hence, LSTD-RP may have a better performance than solving LSTD in $\mathcal{F}$ whenever this additional term is smaller than the gain achieved in the estimation error. Note that $m(\widehat{\Pi}_{\mathcal{F}} v)$ highly depends on the value function V that is being approximated and the features of the space $\mathcal{F}$. It is important to carefully tune the value of d as both the estimation error and the additional approximation error in Eq. 10.1 depend on d. For instance, while a small value of d significantly reduces the estimation error (and the need for samples), it may amplify the additional approximation error term, and thus, reduce the advantage of LSTD-RP over LSTD. We may get an idea on how to select the value of d by optimizing the bound

$$d = \frac{m(\widehat{\Pi}_{\mathcal{F}} v)}{\gamma V_{\max} L} \sqrt{\frac{n \nu_n (1-\gamma)}{1+\gamma}}\,. \tag{10.7}$$

Therefore, when n samples are available the optimal value for d is of the order $O(\sqrt{n})$. Using the value of d in Eq. 10.7, we can rewrite the bound of Eq. 10.1 as (up to the dominated term $1/n$)

$$||v - \widehat{v}||_n \leqslant \frac{1}{\sqrt{1-\gamma^2}}||v - \widehat{\Pi}_{\mathcal{F}} v||_n + \frac{1}{1-\gamma}\sqrt{8\log(8n/\delta)}\sqrt{\gamma V_{\max} L\, m(\widehat{\Pi}_{\mathcal{F}} v)}\Big(\frac{1-\gamma}{n\nu_n(1+\gamma)}\Big)^{1/4}. \tag{10.8}$$

Using Eqs. 10.6 and 10.8, it would be easier to compare the performance of LSTD-RP and LSTD in space $\mathcal{F}$, and observe the role of the term $m(\widehat{\Pi}_{\mathcal{F}} v)$. For further discussion on $m(\widehat{\Pi}_{\mathcal{F}} v)$ refer to Maillard and Munos (2009) and for the case of $D = \infty$ to Section 4.3 of this paper.

Remark 2. As discussed in the introduction, when the dimensionality D of $\mathcal{F}$ is much bigger than the number of samples n, the learning algorithms are likely to overfit the data. In this case, it is reasonable to assume that the target vector v itself belongs to the vector space $\mathcal{F}_n$. We state this condition using the following assumption:

Assumption *(Overfitting).* For any set of n points $\{X_i\}_{i=1}^n$, there exists a function $f \in \mathcal{F}$ such that $f(X_i) = V(X_i),\ 1 \leqslant i \leqslant n$.

Assumption 4.1 is equivalent to require that the rank of the empirical Gram matrices $\frac{1}{n}\Phi^\top\Phi$ to be bigger than n. Note that Assumption 4.1 is likely to hold whenever $D \gg n$, because in this case we can expect that the features to be independent enough on $\{X_i\}_{i=1}^n$ so that the rank of $\frac{1}{n}\Phi^\top\Phi$ to be bigger than n (e.g., if the features are linearly independent on the samples, it is sufficient to have $D \geqslant n$). Under Assumption 4.1 we can remove the empirical approximation error term in Theorem 10.1 and deduce the following result.

Corollary 10.1 *Under Assumption 4.1 and the conditions of Theorem 10.1, with probability $1-\delta$ (w.r.t. the random sample path and the random space), $\widehat{v}$ satisfies*

$$||v - \widehat{v}||_n \leqslant \frac{1}{\sqrt{1-\gamma^2}}\sqrt{\frac{8\log(8n/\delta)}{d}}m(\widehat{\Pi}_{\mathcal{F}}v) + \frac{\gamma V_{\max} L}{1-\gamma}\sqrt{\frac{d}{\nu_n}}\Big(\sqrt{\frac{8\log(4d/\delta)}{n}} + \frac{1}{n}\Big).$$

4.2 Uniqueness of the LSTD-RP Solution

While the results in the previous section hold for any Markov chain, in this section we assume that the Markov chain $\mathcal{M}^\pi$ admits a stationary distribution ρ and is exponentially fast β-mixing with parameters $\bar{\beta}, b, \kappa$, i.e., its β-mixing coefficients satisfy $\beta_i \leqslant \bar{\beta}\exp(-bi^\kappa)$ (see e.g., Sections 8.2 and 8.3 in Lazaric et al. (2010b) for a more detailed definition of β-mixing processes). As shown in Lazaric et al. (2010a,b), if ρ exists, it would be possible to derive a condition for the existence and uniqueness of the LSTD solution depending on the number of samples and the smallest eigenvalue of the Gram matrix defined according to the stationary distribution ρ, i.e., $G \in \mathbb{R}^{D\times D}$, $G_{ij} = \int \varphi_i(x)\varphi_j(x)\rho(dx)$. We now discuss the existence and uniqueness of the LSTD-RP solution. Note that as D increases, the smallest eigenvalue of G is likely to become smaller and smaller. In fact, the more the features in $\mathcal{F}$, the higher the chance for some of them to be correlated under ρ, thus leading to an ill-conditioned matrix G. On the other hand, since $d < D$, the probability that d independent random combinations of φ_i lead to highly correlated features ψ_j is relatively small. In the following we prove that the smallest eigenvalue of the Gram matrix $H \in \mathbb{R}^{d\times d}$, $H_{ij} = \int \psi_i(x)\psi_j(x)\rho(dx)$ in the random space $\mathcal{G}$ is indeed bigger than the smallest eigenvalue of G with high probability.

Lemma 10.2 *Let $\delta > 0$ and $\mathcal{F}$ and $\mathcal{G}$ be linear spaces with dimensions D and d ($d < D$) as defined in Section 2 with $D > d + 2\sqrt{2d\log(2/\delta)} + 2\log(2/\delta)$. Let the elements of the projection matrix A be Gaussian random variables drawn from $\mathcal{N}(0, 1/d)$. Let the Markov chain $\mathcal{M}^\pi$ admit a stationary distribution ρ. Let G and H be the Gram matrices according to ρ for the spaces $\mathcal{F}$ and $\mathcal{G}$, and ω and χ be their smallest eigenvalues. We have with probability $1-\delta$ (w.r.t. the random space)*

$$\chi \geqslant \frac{D}{d}\omega\left(1 - \sqrt{\frac{d}{D}} - \sqrt{\frac{2\log(2/\delta)}{D}}\right)^2. \tag{10.9}$$

Proof: Let $\beta \in \mathbb{R}^d$ be the eigenvector associated to the smallest eigenvalue χ of H, from the definition of the features Ψ of $\mathcal{G}$ ($H = AGA^\top$) and linear algebra, we obtain

$$\chi||\beta||_2^2 = \beta^\top \chi \beta = \beta^\top H \beta = \beta^\top AGA^\top \beta \geqslant \omega ||A^\top \beta||_2^2 = \omega\, \beta^\top AA^\top \beta \geqslant \omega\, \xi\, ||\beta||_2^2\,, \tag{10.10}$$

where ξ is the smallest eigenvalue of the random matrix $AA^\top$, or in other words, $\sqrt{\xi}$ is the smallest singular value of the $D \times d$ random matrix $A^\top$, i.e., $s_{\min}(A^\top) = \sqrt{\xi}$. We now define $B = \sqrt{d}A$. Note that if the elements of A are drawn from the Gaussian distribution $\mathcal{N}(0, 1/d)$, the elements of B are standard Gaussian random variables, and thus, the smallest eigenvalue of $AA^\top$, ξ, can be written as $\xi = s_{\min}^2(B^\top)/d$. There has been extensive work on extreme singular values of random matrices (see e.g., Rudelson and Vershynin (2010)). For a $D \times d$ random matrix with independent standard normal random variables, such as $B^\top$, we have with probability $1 - \delta$ (see Rudelson and Vershynin (2010) for more details)

$$s_{\min}(B^\top) \geqslant \left(\sqrt{D} - \sqrt{d} - \sqrt{2\log(2/\delta)}\right). \tag{10.11}$$

From Eq. 10.11 and the relation between ξ and $s_{\min}(B^\top)$, we obtain

$$\xi \geqslant \frac{D}{d}\left(1 - \sqrt{\frac{d}{D}} - \sqrt{\frac{2\log(2/\delta)}{D}}\right)^2, \tag{10.12}$$

with probability $1 - \delta$. The claim follows by replacing the bound for ξ from Eq. 10.12 in Eq. 10.10. □

The result of Lemma 10.2 is for Gaussian random matrices. However, it would be possible to extend this result using non-asymptotic bounds for the extreme singular values of more general random matrices Rudelson and Vershynin (2010). Note that in Eq. 10.9, D/d is always greater than 1 and the term in the parenthesis approaches 1 for large values of D. Thus, we can conclude that with high probability the smallest eigenvalue χ of the Gram matrix H of the randomly generated low-dimensional space $\mathcal{G}$ is bigger than the smallest eigenvalue ω of the Gram matrix G of the high-dimensional space $\mathcal{F}$.

Lemma 10.3 *Let $\delta > 0$ and $\mathcal{F}$ and $\mathcal{G}$ be linear spaces with dimensions D and d ($d < D$) as defined in Section 2 with $D > d + 2\sqrt{2d\log(2/\delta)} + 2\log(2/\delta)$. Let the elements of the projection matrix A be Gaussian random variables drawn from $\mathcal{N}(0, 1/d)$. Let the Markov chain $\mathcal{M}^\pi$ admit a stationary distribution ρ. Let G be the Gram matrix according to ρ for space $\mathcal{F}$ and ω be its smallest eigenvalue. Let $\{X_t\}_{t=1}^n$ be a trajectory of length n generated by a stationary β-mixing process with stationary distribution ρ. If the number of samples n satisfies*

$$n > \frac{288L^2\, d\, \Lambda(n, d, \delta/2)}{\omega D} \max\left\{\frac{\Lambda(n, d, \delta/2)}{b}, 1\right\}^{1/\kappa} \left(1 - \sqrt{\frac{d}{D}} - \sqrt{\frac{2\log(2/\delta)}{D}}\right)^{-2}, \tag{10.13}$$

where $\Lambda(n, d, \delta) = 2(d+1)\log n + \log\frac{e}{\delta} + \log^+\left(\max\{18(6e)^{2(d+1)}, \bar{\beta}\}\right)$, then with probability $1 - \delta$, the features $\psi_1, \ldots, \psi_d$ are linearly independent on the states $\{X_t\}_{t=1}^n$, i.e., $||g_\beta||_n = 0$ implies $\beta = 0$, and the smallest eigenvalue ν_n of the sample-based Gram matrix $\frac{1}{n}\Psi^\top\Psi$ satisifies

$$\sqrt{\nu_n} \geqslant \sqrt{\nu} = \frac{\sqrt{\omega}}{2}\sqrt{\frac{D}{d}}\left(1 - \sqrt{\frac{d}{D}} - \sqrt{\frac{2\log(\frac{2}{\delta})}{D}}\right) - 6L\sqrt{\frac{2\Lambda(n,d,\frac{\delta}{2})}{n}\max\left\{\frac{\Lambda(n,d,\frac{\delta}{2})}{b},1\right\}^{1/\kappa}} > 0\,. \tag{10.14}$$

Proof: The proof follows similar steps as in Lemma 4 in Lazaric et al. (2010b). A sketch of the proof is available in Ghavamzadeh et al. (2010b). □

By comparing Eq. 10.13 with Eq. 13 in Lazaric et al. (2010b), we can see that the number of samples needed for the empirical Gram matrix $\frac{1}{n}\Psi^\top\Psi$ in $\mathcal{G}$ to be invertible with high probability is less than that for its counterpart $\frac{1}{n}\Phi^\top\Phi$ in the high-dimensional space $\mathcal{F}$.

4.3 Generalization Bound

In this section, we show how Theorem 10.1 can be generalized to the entire state space $\mathcal{X}$ when the Markov chain $\mathcal{M}^\pi$ has a stationary distribution ρ. We consider the case in which the samples $\{X_t\}_{t=1}^n$ are obtained by following a single trajectory in the stationary regime of $\mathcal{M}^\pi$, i.e., when X_1 is drawn from ρ. As discussed in Remark 2 of Section 4.1, it is reasonable to assume that the high-dimensional space $\mathcal{F}$ contains functions that are able to perfectly fit the value function V in any finite number n ($n < D$) of states $\{X_t\}_{t=1}^n$, thus we state the following theorem under Assumption 4.1.

Theorem 10.2 (*Generalization error of LSTD-RP with Markov design*) *Let $\delta > 0$ and $\mathcal{F}$ and $\mathcal{G}$ be linear spaces with dimensions D and d ($d < D$) as defined in Section 2 with $d \geqslant 15\log(8n/\delta)$. Let $\{X_t\}_{t=1}^n$ be a path generated by a stationary β-mixing process with stationary distribution ρ. Let $\widehat{V}$ be the LSTD-RP solution in the random space $\mathcal{G}$. Then under Assumption 4.1, with probability $1-\delta$ (w.r.t. the random sample path and the random space),*

$$||V - T(\widehat{V})||_\rho \leqslant \frac{2}{\sqrt{1-\gamma^2}}\sqrt{\frac{8\log(24n/\delta)}{d}}m(\Pi_{\mathcal{F}}V) + \frac{2\gamma V_{\max}L}{1-\gamma}\sqrt{\frac{d}{\nu}}\left(\sqrt{\frac{8\log(12d/\delta)}{n}} + \frac{1}{n}\right) + \varepsilon\,, \tag{10.15}$$

where ν is a lower bound on the eigenvalues of the Gram matrix $\frac{1}{n}\Psi^\top\Psi$ defined by Eq. 10.14 and

$$\varepsilon = 24V_{\max}\sqrt{\frac{2\Lambda(n,d,\delta/3)}{n}\max\left\{\frac{\Lambda(n,d,\delta/3)}{b},1\right\}^{1/\kappa}}\,.$$

with $\Lambda(n,d,\delta)$ defined as in Lemma 10.3. Note that T in Eq. 10.15 is the truncation operator defined in Section 2.

Proof: The proof is a consequence of applying concentration of measures inequalities for β-mixing processes and linear spaces (see Corollary 18 in Lazaric et al. (2010b)) on the term $||V - T(\widehat{V})||_n$, using the fact that $||V - T(\widehat{V})||_n \leqslant ||V - \widehat{V}||_n$, and using the bound of Corollary 10.1. The bound of Corollary 10.1 and the lower bound on ν, each one holding with

probability $1-\delta'$, thus, the statement of the theorem (Eq. 10.15) holds with probability $1-\delta$ by setting $\delta = 3\delta'$. □

Remark 1. An interesting property of the bound in Theorem 10.2 is that the approximation error of V in space $\mathcal{F}$, $||V - \Pi_{\mathcal{F}} V||_\rho$, does not appear and the error of the LSTD solution in the randomly projected space only depends on the dimensionality d of $\mathcal{G}$ and the number of samples n. However this property is valid only when Assumption 4.1 holds, i.e., at most for $n \leqslant D$. An interesting case here is when the dimension of $\mathcal{F}$ is infinite ($D = \infty$), so that the bound is valid for any number of samples n. In Maillard and Munos (2010a), two approximation spaces $\mathcal{F}$ of infinite dimension were constructed based on a multi-resolution set of features that are rescaled and translated versions of a given mother function. In the case that the mother function is a wavelet, the resulting features, called scrambled wavelets, are linear combinations of wavelets at all scales weighted by Gaussian coefficients. As a results, the corresponding approximation space is a Sobolev space $H^s(\mathcal{X})$ with smoothness of order $s > p/2$, where p is the dimension of the state space $\mathcal{X}$. In this case, for a function $f_\alpha \in H^s(\mathcal{X})$, it is proved that the ℓ_2-norm of the parameter α is equal to the norm of the function in $H^s(\mathcal{X})$, i.e., $||\alpha||_2 = ||f_\alpha||_{H^s(\mathcal{X})}$. We do not describe those results further and refer the interested readers to Maillard and Munos (2010a). What is important about the results of Maillard and Munos (2010a) is that it shows that it is possible to consider infinite dimensional function spaces for which $\sup_x ||\varphi(x)||_2$ is finite and $||\alpha||_2$ is expressed in terms of the norm of f_α in $\mathcal{F}$. In such cases, $m(\Pi_{\mathcal{F}} V)$ is finite and the bound of Theorem 10.2, which does not contain any approximation error of V in $\mathcal{F}$, holds for any n. Nonetheless, further investigation is needed to better understand the role of $||f_\alpha||_{H^s(\mathcal{X})}$ in the final bound.

Remark 2. As discussed in the introduction, regularization methods have been studied in solving high-dimensional RL problems. Therefore, it is interesting to compare our results for LSTD-RP with those reported in Farahmand et al. (2008) for ℓ_2-regularized LSTD. Under Assumption 4.1, when $D = \infty$, by selecting the features as described in the previous remark and optimizing the value of d as in Eq. 10.7, we obtain

$$||V - T(\widehat{V})||_\rho \leqslant O\Big(\sqrt{||f_\alpha||_{H^s(\mathcal{X})}}\, n^{-1/4}\Big). \tag{10.16}$$

Although the setting considered in Farahmand et al. (2008) is different than ours (e.g., the samples are i.i.d.), a qualitative comparison of Eq. 10.16 with the bound in Theorem 2 of Farahmand et al. (2008) shows a striking similarity in the performance of the two algorithms. In fact, they both contain the Sobolev norm of the target function and have a similar dependency on the number of samples with a convergence rate of $O(n^{-1/4})$ (when the smoothness of the Sobolev space in Farahmand et al. (2008) is chosen to be half of the dimensionality of $\mathcal{X}$). This similarity asks for further investigation on the difference between ℓ_2-regularized methods and random projections in terms of prediction performance and computational complexity.

5 LSPI with Random Projections

In this section, we move from policy evaluation to policy iteration and provide a performance bound for LSPI with random projections (LSPI-RP), i.e., a policy iteration algorithm that uses LSTD-RP at each iteration. LSPI-RP starts with an arbitrary initial value function $V_{-1} \in \mathcal{B}(\mathcal{X}; V_{\max})$ and its corresponding greedy policy π_0. At the first iteration, it approximates V^{π_0} using LSTD-RP and returns a function $\widehat{V}_0$, whose truncated version $\tilde{V}_0 = T(\widehat{V}_0)$ is used to build the policy for the second iteration. More precisely, π_1 is a greedy policy w.r.t. $\tilde{V}_0$. So, at each iteration k, a function $\widehat{V}_{k-1}$ is computed as an approximation to $V^{\pi_{k-1}}$, and then truncated, $\tilde{V}_{k-1}$, and used to build the policy π_k.[1] Note that in general, the measure $\sigma \in \mathcal{S}(\mathcal{X})$ used to evaluate the final performance of the LSPI-RP algorithm might be different from the distribution used to generate samples at each iteration. Moreover, the LSTD-RP performance bounds require the samples to be collected by following the policy under evaluation. Thus, we need Assumptions 1-3 in Lazaric et al. (2010b) in order to **1)** define a lower-bounding distribution μ with constant $C < \infty$, **2)** guarantee that with high probability a unique LSTD-RP solution exists at each iteration, and **3)** define the slowest β-mixing process among all the mixing processes $\mathcal{M}^{\pi_k}$ with $0 \leqslant k < K$.

Theorem 10.3 (*Performance bound of LSPI-RP*) *Let $\delta > 0$ and $\mathcal{F}$ and $\mathcal{G}$ be linear spaces with dimensions D and d ($d < D$) as defined in Section 2 with $d \geqslant 15 \log(8Kn/\delta)$. At each iteration k, we generate a path of size n from the stationary β-mixing process with stationary distribution $\rho_{k-1} = \rho^{\pi_{k-1}}$. Let n satisfy the condition in Eq. 10.13 for the slower β-mixing process. Let V_{-1} be an arbitrary initial value function, $\widehat{V}_0, \ldots, \widehat{V}_{K-1}$ ($\tilde{V}_0, \ldots, \tilde{V}_{K-1}$) be the sequence of value functions (truncated value functions) generated by LSPI-RP, and π_K be the greedy policy w.r.t. $\tilde{V}_{K-1}$. Then, under Assumption 4.1 and Assumptions 1-3 in Lazaric et al. (2010b), with probability $1-\delta$ (w.r.t. the random samples and the random spaces), we have*

$$||V^* - V^{\pi_K}||_\sigma \leqslant \frac{4\gamma}{(1-\gamma)^2}\left\{(1+\gamma)\sqrt{CC_{\sigma,\mu}}\left[\frac{2V_{\max}}{\sqrt{1-\gamma^2}}\sqrt{\frac{C}{\omega_\mu}}\sqrt{\frac{8\log(24Kn/\delta)}{d}}\sup_{x\in\mathcal{X}}||\varphi(x)||_2\right.\right. \quad (10.17)$$
$$\left.\left. + \frac{2\gamma V_{\max} L}{1-\gamma}\sqrt{\frac{d}{\nu_\mu}}\left(\sqrt{\frac{8\log(12Kd/\delta)}{n}} + \frac{1}{n}\right) + \mathcal{E}\right] + \gamma^{\frac{K-1}{2}} R_{\max}\right\},$$

where $C_{\sigma,\mu}$ is the concentrability term from Definition 2 in Antos et al. (2008), ω_μ is the smallest eigenvalue of the Gram matrix of space $\mathcal{F}$ w.r.t. μ, ν_μ is ν from Eq. 10.14 in which ω is replaced by ω_μ, and $\mathcal{E}$ is ε from Theorem 10.2 written for the slowest β-mixing process.

Proof: The proof follows similar lines as in the proof of Thm. 8 in Lazaric et al. (2010b) and is available in Ghavamzadeh et al. (2010b). □

[1] Note that the MDP model is needed to generate a greedy policy π_k. In order to avoid the need for the model, we can simply move to LSTD-Q with random projections. Although the analysis of LSTD-RP can be extended to action-value functions and LSTD-RP-Q, for simplicity we use value functions in the following.

Remark. The most critical issue about Theorem 10.3 is the validity of Assumptions 1-3 in Lazaric et al. (2010b). It is important to note that Assumption 1 is needed to bound the performance of LSPI independent from the use of random projections (see Lazaric et al. (2010b)). On the other hand, Assumption 2 is explicitly related to random projections and allows us to bound the term $m(\Pi_{\mathcal{F}}V)$. In order for this assumption to hold, the features $\{\varphi_j\}_{j=1}^{D}$ of the high-dimensional space $\mathcal{F}$ should be carefully chosen so as to be linearly independent w.r.t. μ.

6 Conclusion

Learning in high-dimensional linear spaces is particularly appealing in RL because it allows to have a very accurate approximation of value functions. Nonetheless, the larger the space, the higher the need of samples and the risk of overfitting. In this paper, we introduced an algorithm, called LSTD-RP, in which LSTD is run in a low-dimensional space obtained by a random projection of the original high-dimensional space. We theoretically analyzed the performance of LSTD-RP and showed that it solves the problem of overfitting (i.e., the estimation error depends on the value of the low dimension) at the cost of a slight worsening in the approximation accuracy compared to the high-dimensional space. We also analyzed the performance of LSPI-RP, a policy iteration algorithm that uses LSTD-RP for policy evaluation. The analysis reported in the paper opens a number of interesting research directions such as: **1)** comparison of LSTD-RP to ℓ_2 and ℓ_1 regularized approaches, and **2)** a thorough analysis of the case when $D = \infty$ and the role of $||f_\alpha||_{H^s(\mathcal{X})}$ in the bound.

7 Technical details

7.1 Uniqueness of the LSTD-RP Solution (Proof of Lemma 3)

Proof: [Proof of Lemma 3 - Sketch] Following similar steps as in Lemma 4 in Lazaric et al. (2010b) and using Lemma 2, for any $\beta \in \mathbb{R}^d$ with probability $1 - \delta' - \delta''$ we obtain

$$2||g_\beta||_n + \varepsilon \geqslant \sqrt{\chi}||\beta||_2 \geqslant ||\beta||_2 \sqrt{\frac{D\,\omega}{d}} \left(1 - \sqrt{\frac{d}{D}} - \sqrt{\frac{2\log(2/\delta')}{D}}\right), \tag{10.18}$$

where

$$\varepsilon = 12L||\beta||_2 \sqrt{\frac{2\Lambda(n,d,\delta'')}{n}} \max\left\{\frac{\Lambda(n,d,\delta'')}{b}, 1\right\}^{1/\kappa}. \tag{10.19}$$

Let β be such that $||g_\beta||_n = 0$, then if the number of samples n satisfies the condition in Lemma 3, we may deduce from Eq. 10.18 and 10.19 that $\beta = 0$. This indicates that given the number of samples from Lemma 3, with probability $1-\delta''$, the features $\Psi_1, \ldots, \Psi_d$ are linearly independent on the states $\{X_t\}_{t=1}^n$, and thus, $\nu_n > 0$. The second statement of the lemma is obtained by choosing β to be the eigenvector of the Gram matrix $\frac{1}{n}\Psi^\top\Psi$ corresponding

to the smallest eigenvalue ν_n. For this value of β, we have $||g_\beta||_n = \sqrt{\nu_n}||\beta||$. Finally, both statements of the lemma are obtained by setting $\delta' = \delta'' = \delta/2$. □

7.2 LSPI with Random Projections (Proof of Theorem 3)

We report Assumptions 1-3 in Lazaric et al. (2010b).

Assumption *(Lower-bounding distribution)*There exist a distribution $\mu \in \mathcal{S}(\mathcal{X})$ such that for any policy π that is greedy w.r.t. a function in the truncated space $\tilde{\mathcal{F}}$, $\mu \leqslant C\rho^\pi$, where $C < \infty$ is a constant and ρ^π is the stationary distribution of policy π. Furthermore, given the target distribution $\sigma \in \mathcal{S}(\mathcal{X})$, we assume $C_{\sigma,\mu} < \infty$, where $C_{\sigma,\mu}$ is the concentrability term from Definition 2 in Antos et al. (2008).

Assumption *(Linear independent features)* Let $\mu \in \mathcal{S}(\mathcal{X})$ be the lower-bounding distribution from Assumption 7.2. We assume that the features $\varphi(\cdot)$ of the function space $\mathcal{F}$ are linearly independent w.r.t. μ. In this case, the smallest eigenvalue ω_μ of the Gram matrix $G_\mu \in \mathbb{R}^{D\times D}$ w.r.t. μ is strictly positive.

Assumption *(Slower β-mixing process)* We assume that there exists a stationary β-mixing process with parameters $\bar{\beta}, b, \kappa$, such that for any policy π that is greedy w.r.t. a function in the truncated space $\tilde{\mathcal{F}}$, it is slower than the stationary β-mixing process with stationary distribution ρ^π (with parameters $\bar{\beta}_\pi, b_\pi, \kappa_\pi$). This means that $\bar{\beta}$ is larger and b and κ are smaller than their counterparts $\bar{\beta}_\pi$, b_π, and κ_π.

Proof: We first notice that the equality

$$(I - \gamma P^{\pi_k})(\tilde{V}_k - V^{\pi_k}) = \tilde{V}_k - \mathcal{T}^{\pi_k}\tilde{V}_k$$

holds component-wise for any $x \in \mathcal{X}$. Let $\varepsilon_k = (\tilde{V}_k - V^{\pi_k})$ and ρ_k be the stationary distribution of π_k. We have

$$||\tilde{V}_k - \mathcal{T}^{\pi_k}\tilde{V}_k||_{\rho_k} = ||(I - \gamma P^{\pi_k})\varepsilon_k||_{\rho_k} \leqslant (1 + \gamma||P^{\pi_k}||_{\rho_k})||\varepsilon_k||_{\rho_k} = (1+\gamma)||\varepsilon_k||_{\rho_k},$$

where we used the fact that P^{π_k} is the transition kernel for policy π_k and ρ_k is its stationary distribution. From a direct application of Lemma 13 in Munos and Szepesvári (2008) and the previous inequality, after K iterations we obtain

$$||V^* - V^{\pi_K}||_\sigma \leqslant \frac{4\gamma}{(1-\gamma)^2}\left[(1+\gamma)\max_{0\leqslant k<K} C_{\sigma,\rho_k}^{1/2}||\varepsilon_k||_{\rho_k} + \gamma^{\frac{K-1}{2}}R_{\max}\right],$$

where $||\varepsilon_k||_{\rho_k}$ is bounded as in Theorem 2 in the paper. The main issue in the previous bound is the maximization over the iterations. We first focus on the maximum of the error $||\varepsilon_k||_{\rho_k}$. The only term in the statement of Theorem 2 explicitly depending on the specific iteration is the magnitude (notice that the target function at iteration k is V^{π_k})

$$\max_{0\leqslant k<K} m\big(\Pi_{\mathcal{F}}^k(V^{\pi_k})\big) = \max_{0\leqslant k<K}||\alpha_k||_2 \sup_{x\in\mathcal{X}}||\varphi(x)||_2\,, \tag{10.20}$$

where $\Pi^k_{\mathcal{F}}$ is the projection operator onto space $\mathcal{F}$ w.r.t. ρ_k and $f_{\alpha_k} = \Pi^k_{\mathcal{F}}(V^{\pi_k})$. Since α_k is a random variable we cannot bound its ℓ_2-norm directly. Under Assumptions 2 and 3 it is possible to show that for any $f_\alpha \in \mathcal{F}$

$$C||f_\alpha||^2_{\rho_k} \geqslant ||f_\alpha||^2_\mu = ||\varphi(\cdot)\alpha||^2_\mu = \alpha^\top G_\mu \alpha \geqslant \omega_\mu ||\alpha||^2_2 \,,$$

where G_μ is the Gram matrix computed w.r.t. distribution μ and ω_μ is its smallest eigenvalue. Here we used the fact that under Assumption 3, G_μ is full rank. As a result, for any iteration k we have

$$||\alpha_k||^2_2 \leqslant \frac{1}{\omega_\mu}||f_{\alpha_k}||^2_\mu \leqslant \frac{C}{\omega_\mu}||f_{\alpha_k}||^2_{\rho_k} \leqslant \frac{C}{\omega_\mu}||f_{\alpha_k}||^2_\infty \,.$$

The function f_{α_k} is the result of a projection w.r.t. norm ρ_k of V^{π_k}, which is bounded by $V_{\max}$. Since $\Pi^k_{\mathcal{F}}$ is a non-expansion in ρ_k norm, $||f_{\alpha_k}||_\infty$ is upper bounded by $V_{\max}$. Thus, the term in Eq. 10.20 can be bounded by [2]

$$\max_{0 \leqslant k < K} m(\Pi^k_{\mathcal{F}}(V^{\pi_k})) \leqslant \sqrt{\frac{C}{\omega_\mu}} V_{\max} \sup_{x \in \mathcal{X}} ||\varphi(x)||_2 \,.$$

Now we bound the concentrability term C_{σ,ρ_k}. From the definition of the concentrability term and Assumption 2, we obtain

$$c_{\sigma,\rho_k}(m) = \sup_{\pi_1,\dots,\pi_s} \left|\left| \frac{d(\mu P^{\pi_1} P^{\pi_1} \dots P^{\pi_s})}{d\rho_k} \right|\right| \leqslant C \sup_{\pi_1,\dots,\pi_s} \left|\left| \frac{d(\mu P^{\pi_1} P^{\pi_1} \dots P^{\pi_s})}{d\mu} \right|\right| = C \cdot c_{\sigma,\mu}(m).$$

Thus, $C_{\sigma,\rho_k} \leqslant C \cdot C_{\sigma,\mu}$. Putting everything together and reordering we obtain the final statement. Finally, we discuss about the eigenvalues of the sequence of Gram matrices G_{ρ_k} obtained through iterations. By Assumption 2 and the definition of G_μ we have

$$(G_\mu)_{ij} = \int_{\mathcal{X}} \varphi_i(x)\varphi_j(x)\mu(dx) \leqslant C \int_{\mathcal{X}} \varphi_i(x)\varphi_j(x)\rho_k(dx) = C(G_{\rho_k})_{ij} \,.$$

Let ω_μ be the smallest eigenvalue of G_μ, ω_k be the smallest eigenvalue of G_{ρ_k}, and α be the eigenvector corresponding to ω_k. We have

$$\omega_\mu ||\alpha||^2_2 \leqslant \alpha^\top G_\mu \alpha \leqslant C\alpha^\top G_{\rho_k}\alpha = C\omega_k \alpha^\top \alpha = C\omega_k ||\alpha||^2_2 \,,$$

thus, obtaining $\omega_\mu \leqslant C\omega_k$. □

[2] Note that the remaining term $\sup_{x \in \mathcal{X}} ||\varphi(x)||_2$ does not depend on k and its specific value depends on the feature space $\varphi(\cdot)$ of $\mathcal{F}$.

Chapter 11

Selecting the State-Representation in Reinforcement Learning.

In this chapter, we consider the problem of selecting the right state-representation in a reinforcement learning problem. Several models (functions mapping past observations to a finite set) of the observations are given, and it is known that for at least one of these models the resulting state dynamics are indeed Markovian. Without knowing neither which of the models is the correct one, nor what are the probabilistic characteristics of the resulting MDP, it is required to obtain as much reward as the optimal policy for the correct model (or for the best of the correct models, if there are several). We propose an algorithm that achieves that, with a regret of order $T^{2/3}$ where T is the horizon time.

The work presented in this chapter is a joint work with *Daniil Ryabko* and has been accepted to the 25*th conference on advances in Neural Information Processing Systems (NIPS 2011)*.

Contents

1 Introduction

We consider the problem of selecting the right state-representation in an average-reward reinforcement learning problem. Each state-representation is defined by a model φ_j (to which

corresponds a state space $\mathcal{S}_{\varphi_j}$) and we assume that the number J of available models is finite and that (at least) one model is a weakly-communicating Markov decision process (MDP). We do not make any assumption at all about the other models. This problem is considered in the general reinforcement learning setting, where an agent interacts with an unknown environment in a single stream of repeated observations, actions and rewards. There are no "resets," thus all the learning has to be done online. Our goal is to construct an algorithm that performs almost as well as the algorithm that knows both which model is a MDP (knows the "true" model) and the characteristics of this MDP (the transition probabilities and rewards).

Consider some examples that help motivate the problem. The first example is high-level feature selection. Suppose that the space of histories is huge, such as the space of video streams or that of game plays. In addition to these data, we also have some high-level features extracted from it, such as "there is a person present in the video" or "the adversary (in a game) is aggressive." We know that most of the features are redundant, but we also know that some combination of some of the features describes the problem well and exhibits Markovian dynamics. Given a potentially large number of feature combinations of this kind, we want to find a policy whose average reward is as good as that of the best policy for the right combination of features. Another example is bounding the order of an MDP. The process is known to be k-order Markov, where k is unknown but un upper bound $K >> k$ is given. The goal is to perform as well as if we knew k. Yet another example is selecting the right discretization. The environment is an MDP with a continuous state space. We have several candidate quantizations of the state space, one of which gives an MDP. Again, we would like to find a policy that is as good as the optimal policy for the right discretization. This example also opens the way for extensions of the proposed approach: we would like to be able to treat an infinite set of possible discretization, none of which may be perfectly Markovian. The present work can be considered the first step in this direction.

It is important to note that we do not make any assumptions on the "wrong" models (those that do not have Markovian dynamics). Therefore, we are not able to *test* which model is Markovian in the classical statistical sense, since in order to do that we would need a viable alternative hypothesis (such as, the model is not Markov but is K-order Markov). In fact, the constructed algorithm never "knows" which model is the right one; it is "only" able to get the same average level of reward as if it knew.

Previous work. This work builds on previous work on learning average-reward MDPs. Namely, we use in our algorithm as a subroutine the algorithm UCRL2 of Jaksch et al. (2010) that is designed to provide finite time bounds for undiscounted MDPs. Such a problem has been pioneered in the reinforcement learning literature by Kearns and Singh (2002) and then improved in various ways by Brafman and Tennenholtz (2003), Strehl et al. (2006), Tewari and Bartlett (2007), Jaksch et al. (2010), Bartlett and Tewari (2009); UCRL2 achieves a regret of the order $DT^{1/2}$ in any weakly-communicating MDP with diameter D, with respect to the best policy for this MDP. The diameter D of a MDP is defined in Jaksch et al. (2010) as the expected minimum time required to reach any state starting from any other state. A related result is reported in Bartlett and Tewari (2009), which improves on constants related

to the characteristics of the MDP.

A similar approach has been considered in Ryabko and Hutter (2008); the difference is that in that work the probabilistic characteristics of each model are completely known, but the models are not assumed to be Markovian, and belong to a countably infinite (rather than finite) set.

The problem we address can be also viewed as a generalization of the bandit problem (see e.g. Robbins (1952), Lai and Robbins (1985), Auer et al. (2002)): there are finitely many "arms", corresponding to the policies used in each model, and one of the arms is the best, in the sense that the corresponding model is the "true" one. In the usual bandit setting, the rewards are assumed to be i.i.d. thus one can estimate the mean value of the arms while switching arbitrarily from one arm to the next (the quality of the estimate only depends on the number of pulls of each arm). However, in our setting, estimating the average-reward of a policy requires playing it *many times consecutively*. This can be seen as a bandit problem with dependent arms, with complex costs of switching between arms.

Contribution. We show that despite the fact that the true Markov model of states is unknown and that nothing is assumed on the wrong representations, it is still possible to derive a finite-time analysis of the regret for this problem. This is stated in Theorem 11.1; the bound on the regret that we obtain is of order $T^{2/3}$.

The intuition is that if the "true" model φ^* is known, but its probabilistic properties are not, then we still know that there exists an optimal control policy that depends on the observed state $s_{j^*,t}$ only. Therefore, the optimal rate of rewards can be obtained by a clever exploration/exploitation strategy, such as UCRL2 algorithm Jaksch et al. (2010). Since we do not know in advance which model is a MDP, we need to explore them all, for a sufficiently long time in order to estimate the rate of rewards that one can get using a good policy in that model.

Outline. In Section 2 we introduce the precise notion of model and set up the notations. Then we present the proposed algorithm in Section 3; it uses UCRL2 of Jaksch et al. (2010) as a subroutine and selects the models φ according to a penalized empirical criterion. In Section 4 we discuss some directions for further development. Finally, Section 5 is devoted to the proof of Theorem 11.1.

2 Notation and definitions

We consider a space of observations $\mathcal{O}$, a space of actions $\mathcal{A}$, and a space of rewards $\mathcal{R}$ (all assumed to be Polish). Moreover, we assume that $\mathcal{A}$ is of finite cardinality $A \stackrel{\text{def}}{=} |\mathcal{A}|$ and that $0 \in \mathcal{R} \subset [0,1]$. The set of histories up to time t for all $t \in \mathbb{N} \cup \{0\}$ will be denoted by $\mathcal{H}_{<t} \stackrel{\text{def}}{=} \mathcal{O} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{O})^{t-1}$, and we define the set of all possible histories by $\mathcal{H} \stackrel{\text{def}}{=} \bigcup_{t=1}^{\infty} \mathcal{H}_{<t}$.

Environments. For a Polish $\mathcal{X}$, we Denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$. Define an environment to be a mapping from the set of histories $\mathcal{H}$ to the set of functions that map any action $a \in \mathcal{A}$ to a probability distribution $\nu_a \in \mathcal{P}(\mathcal{R} \times \mathcal{O})$ over the product space of rewards and observations.

We consider the problem of reinforcement learning when the learner interacts with some *unknown* environment $e^\star$. The interaction is sequential and goes as follows: first some $h_{<1} = \{o_0\}$ is generated according to ι, then at time step $t > 0$, the learner choses an action $a_t \in \mathcal{A}$ according to the current history $h_{<t} \in \mathcal{H}_{<t}$. Then a couple of reward and observations (r_t, o_t) is drawn according to the distribution $(e^\star(h_{<t}))_{a_t} \in \mathcal{P}(\mathcal{R} \times \mathcal{O})$. Finally, $h_{<t+1}$ is defined by the concatenation of $h_{<t}$ with (a_t, r_t, o_t). With these notations, at each time step $t > 0$, o_{t-1} is the last observation given to the learner before choosing an action, a_t is the action output at this step, and r_t is the immediate reward received after playing a_t.

State representation functions (models). Let $\mathcal{S} \subset \mathbb{N}$ be some finite set; intuitively, this has to be considered as a set of states. A *state representation* function φ is a function from the set of histories $\mathcal{H}$ to $\mathcal{S}$. For a state representation function φ, we will use the notation $\mathcal{S}_\varphi$ for its set of states, and $s_{t,\varphi} := \varphi(h_{<t})$.

In the sequel, when we talk about a Markov decision process, it will be assumed to be *weakly communicating*, which means that for each pair of states u_1, u_2 there exists $k \in \mathbb{N}$ and a sequence of actions $\alpha_1, .., \alpha_k \in \mathcal{A}$ such that $P(s_{k+1,\varphi} = u_2 | s_{1,\varphi} = u_1, a_1 = \alpha_1 ... a_k = \alpha_k) > 0$. Having that in mind, we introduce the following definition.

Definition 11.1 *We say that an environment e with a state representation function φ is Markov, or, for short, that φ is a Markov model (of e), if the process $(s_{t,\varphi}, a_t, r_t), t \in \mathbb{N}$ is a (weakly communicating) Markov decision process.*

For example, consider a state-representation function φ that depends only on the last observation, and that partitions the observation space into finitely many cells. Then an environment is Markov with this representation function if the probability distribution on the next cells only depends on the last observed cell and action. Note that there may be many state-representation functions with which an environment e is Markov.

3 Main results

Given a set $\Phi = \{\varphi_j;\, j \leqslant J\}$ of J state-representation functions (models), one of which being a Markov model of the unknown environment $e^\star$, we want to construct a strategy that performs nearly as well as the best algorithm that knows which φ_j is Markov, and knows all the probabilistic characteristics (transition probabilities and rewards) of the MDP corresponding to this model. For that purpose we define the regret of any strategy at time T, like in Jaksch et al. (2010), Bartlett and Tewari (2009), as

$$\Delta(T) \stackrel{\text{def}}{=} T\rho^\star - \sum_{t=1}^{T} r_t \,,$$

where r_t are the rewards received when following the proposed strategy and $\rho^\star$ is the average optimal value in the best Markov model, i.e., $\rho^\star = \lim_T \frac{1}{T}\mathbb{E}(\sum_{t=1}^{T} r_t(\pi^\star))$ where $r_t(\pi^\star)$ are the rewards received when following the optimal policy for the best Markov model. Note that this definition makes sense since when the MDP is weakly communicating, the average optimal value of reward does not depend on the initial state. Also, one could replace

$T\rho^*$ with the expected sum of rewards obtained in T steps (following the optimal policy) at the price of an additional $O(\sqrt{T})$ term.

In the next subsection, we describe an algorithm that achieves a sub-linear regret of order $T^{2/3}$.

3.1 Best Lower Bound (BLB) algorithm

In this section, we introduce the Best-Lower-Bound (BLB) algorithm, described in Figure 11.1.

The algorithm works in stages of doubling length. Each stage consists in 2 phases: an exploration and an exploitation phase. In the exploration phase, BLB plays the UCRL2 algorithm on each model $(\varphi_j)_{1\leqslant j\leqslant J}$ successively, as if each model φ_j was a Markov model, for a fixed number $\tau_{i,1,J}$ of rounds. The exploitation part consists in selecting first the model with highest lower bound, according to the empirical rewards obtained in the previous exploration phase. This model is initially selected for the same time as in the exploration phase, and then a test decides to either continue playing this model (if its performance during exploitation is still above the corresponding lower bound, i.e. if the rewards obtained are still at least as good as if it was playing the best model). If it does not pass the test, then another model (with second best lower-bound) is select and played, and so on. Until the exploitation phase (of fixed length $\tau_{i,2}$) finishes and the next stage starts.

The length of stage i is fixed and defined to be $\tau_i \stackrel{\text{def}}{=} 2^i$. Thus for a total time horizon T, the number of stages $I(T)$ before time T is $I(T) \stackrel{\text{def}}{=} \llcorner \log_2(T+1) \lrcorner$. Each stage i (of length τ_i) is further decomposed into an exploration (length $\tau_{i,1}$) and an exploitation (length $\tau_{i,2}$) phases.

Exploration phase. All the models $\{\varphi_j\}_{j\leqslant J}$ are played one after another for the same amount of time $\tau_{i,1,J} \stackrel{\text{def}}{=} \frac{\tau_{i,1}}{J}$. Each episode $1 \leqslant j \leqslant J$ consists in running the UCRL2 algorithm using the model of states and transitions induced by the state-representation function φ_j. Note that UCRL2 does not require the horizon T in advance, but requires a parameter p in order to ensure a near optimal regret bound with probability higher than $1-p$. We define this parameter p to be $\delta_i(\delta)$ in stage i, where

$$\delta_i(\delta) \stackrel{\text{def}}{=} (2^i - (J^{-1}+1)2^{2i/3} + 4)^{-1} 2^{-i+1}\delta\,. \tag{11.1}$$

The average empirical reward received during each episode is written $\widehat{\mu}_{i,1}(\varphi_j)$.

Exploitation phase. We use the empirical rewards $\widehat{\mu}_{i,1}(\varphi_j)$ received in the previous exploration part of stage i together with a confidence bound in order to select the model to play. Moreover, a model φ is no longer run for a fixed period of time (as in the exploration part of stage i), but for a period $\tau_{i,2}(\varphi)$ that depends on some test; we first initialize $\mathcal{J} := \{1,\dots,J\}$ and then choose

$$\widehat{j} \stackrel{\text{def}}{=} \arg\max_{j\in\mathcal{J}} \widehat{\mu}_{i,1}(\varphi_j) - 2B(i,\varphi_j,\delta)\,, \tag{11.2}$$

Parameters: f, δ
For each stage $i \geqslant 1$ do
Set the total length of stage i to be $\tau_i := 2^i$.

1. Exploration. Set $\tau_{i,1} = \tau_i^{2/3}$. For each $j \in \{1, \ldots, J\}$ do
   - Run UCRL2 with parameter $\delta_i(\delta)$ defined in (11.1) using φ_j during $\tau_{i,1,J}$ steps: the state space is assumed to be $\mathcal{S}_{\varphi_j}$ with transition structure induced by φ_j.
   - Compute the corresponding average empirical reward $\widehat{\mu}_{i,1}(\varphi_j)$ received during this exploration phase.
2. Exploitation. Set $\tau_{i,2} = \tau_i - \tau_{i,1}$ and initialize $\mathcal{J} := \{1, \ldots, J\}$.
   While the current length of the exploitation part is less than $\tau_{i,2}$ do
   - Select $\widehat{j} = \arg\max_{j \in \mathcal{J}} \widehat{\mu}_{i,1}(\varphi_j) - 2B(i, \varphi_j, \delta)$ (using (11.3)).
   - Run UCRL2 with parameter $\delta_i(\delta)$ using $\varphi_{\widehat{j}}$: update at each time step t the current average empirical reward $\widehat{\mu}_{i,2,t}(\varphi_{\widehat{j}})$ from the beginning of the run. Provided that the length of the current run is larger than $\tau_{i,1,J}$, do the test
   $$\widehat{\mu}_{i,2,t}(\varphi_{\widehat{j}}) \geqslant \widehat{\mu}_{i,1}(\varphi_{\widehat{j}}) - 2B(i, \varphi_{\widehat{j}}, \delta)\,.$$
   - If the test fails, then stop UCRL2 and set $\mathcal{J} := \mathcal{J} \setminus \{\widehat{j}\}$. If $\mathcal{J} = \varnothing$ then set $J := \{1, \ldots, J\}$.

Figure 11.1: The Best-Lower-Bound selection strategy.

where we define

$$B(i,\varphi,\delta) \stackrel{\text{def}}{=} 34 f(\tau_i - 1 + \tau_{i,1})|\mathcal{S}_\varphi| \sqrt{\frac{A \log(\frac{\tau_{i,1,J}}{\delta_i(\delta)})}{\tau_{i,1,J}}}, \tag{11.3}$$

where δ and the function f are parameters of the BLB algorithm. Then UCRL2 is played using the selected model $\varphi_{\widehat{j}}$ for the parameter $\delta_i(\delta)$. In parallel we test whether the average empirical reward we receive during this exploitation phase is high enough; at time t, if the length of the current episode is larger than $\tau_{1,i,J}$, we test if

$$\widehat{\mu}_{i,2,t}(\varphi_{\widehat{j}}) \geqslant \widehat{\mu}_{i,1}(\varphi_{\widehat{j}}) - 2B(i,\varphi_{\widehat{j}},\delta). \tag{11.4}$$

If the test is positive, we keep playing UCRL2 using the same model. Now, if the test fails, then the model $\widehat{j}$ is discarded (until the end of stage i) i.e. we update $\mathcal{J} := \mathcal{J} \setminus \{\widehat{j}\}$ and we select a new one according to (11.2). We repeat those steps until the total time $\tau_{i,2}$ of the exploitation phase of stage i is over.

Remark Note that the model selected for exploitation in (11.2) is the one that has the best lower bound. This is a pessimistic (or robust) selection strategy. We know that if the right model is selected, then with high probability, this model will be kept during the whole exploitation phase. If this is not the right model, then either the policy provides good rewards and we should keep playing it, or it does not, in which case it will not pass the test (11.4) and will be removed from the set of models that will be exploited in this phase.

3.2 Regret analysis

Theorem 11.1 (Main result) *Assume that a finite set of J state-representation functions Φ is given, and there exists at least one function $\varphi^\star \in \Phi$ such that with $\varphi^\star$ as a state-representation function the environment is a Markov decision process. If there are several such models, let $\varphi^\star$ be the one with the highest average reward of the optimal policy of the corresponding MDP. Then the regret (with respect to the optimal policy corresponding to φ^*) of the BLB algorithm run with parameter δ, for any horizon T, with probability higher than $1-\delta$ is bounded as follows*

$$\Delta(T) \leqslant c f(T) S \Big(AJ \log\left((J\delta)^{-1}\right) \log_2(T) \Big)^{1/2} T^{2/3} + c' DS \Big(A \log(\delta^{-1}) \log_2(T) T \Big)^{1/2} + c(f,D), \tag{11.5}$$

for some numerical constants c, c' and $c(f,D)$. The parameter $f(t)$ can be chosen to be any increasing function, for instance the choice $f(t) := \log_2 t + 1$, gives $c(f,D) \leqslant 2^D$.

The proof of this result is reported in Section 5.

Remark. Importantly, the algorithm considered here *does not* know in advance the diameter D of the true model, nor the time horizon T. Due to this lack of knowledge, it uses a guess $f(t)$ (e.g. $\log(t)$) on this diameter, which result in the additional regret term $c(f,D)$ and the additional factor $f(T)$; knowing D would enable to remove both of them, but this is a strong assumption. Choosing $f(t) := \log_2 t + 1$ gives a bound which is of order $T^{2/3}$ in T but is exponential in D; taking $f(t) := t^\varepsilon$ we get a bound of order $T^{2/3+\varepsilon}$ in T but of polynomial order $1/\varepsilon$ in D.

4 Discussion and outlook

Intuition. The main idea why this algorithm works is as follows. The "wrong" models are used during exploitation stages only as long as they are giving rewards that are higher than the rewards that could be obtained in the "true" model. All the models are explored sufficiently long so as to be able to estimate the optimal reward level in the true model, and to learn its policy. Thus, nothing has to be known about the "wrong" models. This is in stark contrast to the usual situation in mathematical statistics, where to be able to test a hypothesis about a model (e.g., that the data is generated by a certain model versus some alternative models), one has to make assumptions about alternative models. This has to be done in order to make sure that the Type II error is small (the power of the test is large): that this error is small has to be proven under the alternative. Here, although we are solving seemingly the same problem, the role of the Type II error is played by the rewards. As long as the rewards are high we do not care where the model we are using is correct or not. We only have to ensure that the true model passes the test.

Assumptions. A crucial assumption made in this work is that the "true" model φ^* belongs to a known finite set. While passing from a finite to a countably infinite set appears rather straightforward, getting rid of the assumption that this set *contains* the true model seems more difficult. What one would want to obtain in this setting is sub-linear regret with respect to the performance of the optimal policy in the best model; this, however, seems difficult without additional assumptions on the probabilistic characteristics of the models. Another approach not discussed here would be to try to *build* a good state representation function, as what is suggested for instance in Hutter (2009). Yet another interesting generalization in this direction would be to consider uncountable (possibly parametric but general) sets of models. This, however, would necessarily require some heavy assumptions on the set of models.

Regret. The reader familiar with adversarial bandit literature will notice that our bound of order $T^{2/3}$ is worse than $T^{1/2}$ that usually appears in this context (see, for example Auer et al. (1995)). The reason is that our notion of regret is different: in adversarial bandit literature, the regret is measured with respect to the best choice of the arm *for the given fixed history*. In contrast, we measure the regret with respect to the best policy (for knows the correct model and its parameters) that, in general, would obtain completely different (from what our algorithm would get) rewards and observations right from the beginning.

Estimating the diameter? As previously mentioned, a possibly large additive constant $c(f, D)$ appears in the regret since we do not known a bound on the diameter of the MDP in the "true" model, and use $\log T$ instead. Finding a way to properly address this problem by estimating online the diameter of the MDP is an interesting open question. Let us provide some intuition concerning this problem. First, we notice that, as reported in Jaksch et al. (2010), when we compute an optimistic model based on the empirical rewards and transitions of the true model, the span of the corresponding optimistic value function $sp(\widehat{V}^+)$ is always smaller than the diameter D. This span increases as we get more rewards and transitions

samples, which gives a natural empirical lower bound on D. However, it seems quite difficult to compute a tight empirical upper bound on D (or $sp(\widehat{V}^+)$). In Bartlett and Tewari (2009), the authors derive a regret bound that scales with the span of the true value function $sp(V^\star)$, which is also less than D, and can be significantly smaller in some cases. However, since we do not have the property that $sp(\widehat{V}^+) \leqslant sp(V^\star)$, we need to introduce an explicit penalization in order to control the span of the computed optimistic models, and this requires assuming we know an upper bound B on $sp(V^\star)$ in order to guarantee a final regret bound scaling with B. Unfortunately this does not solve the estimation problem of D, which remains an open question.

5 Proof of Theorem 11.1

In this section, we now detail the proof of Theorem 11.1. The proof is stated in several parts. First we remind a general confidence bound for the UCRL2 algorithm in the true model. Then we decompose the regret into the sum of the regret in each stage i. After analyzing the contribution to the regret in stage i, we then gather all stages and tune the length of each stage and episode in order to get the final regret bound.

5.1 Upper and Lower confidence bounds

From the analysis of UCRL2 in Jaksch et al. (2010), we have the property that with probability higher than $1-\delta'$, the regret of UCRL2 when run for τ consecutive many steps from time t_1 in the true model $\varphi^\star$ is upper bounded by

$$\rho^\star - \frac{1}{\tau}\sum_{t=t_1}^{t_1+\tau-1} r_t \leqslant 34D|\mathcal{S}_{\varphi^\star}|\sqrt{\frac{A\log(\frac{\tau}{\delta'})}{\tau}}\,, \tag{11.6}$$

where D is the diameter of the MDP. What is interesting is that this diameter does not need to be known by the algorithm. Also by carefully looking at the proof of UCRL, it can be shown that the following bound is also valid with probability higher than $1-\delta'$:

$$\frac{1}{\tau}\sum_{t=t_1}^{t_1+\tau-1} r_t - \rho^\star \leqslant 34D|\mathcal{S}_{\varphi^\star}|\sqrt{\frac{A\log(\frac{\tau}{\delta'})}{\tau}}\,.$$

We now define the following quantity, for every model φ, episode length τ and $\delta' \in (0,1)$

$$B_D(\tau,\varphi,\delta') \stackrel{\text{def}}{=} 34D|\mathcal{S}_{\varphi}|\sqrt{\frac{A\log(\frac{\tau}{\delta'})}{\tau}}\,. \tag{11.7}$$

5.2 Regret of stage i

In this section we analyze the regret of the stage i, which we denote Δ_i. Note that since each stage $i \leqslant I$ is of length $\tau_i = 2^i$ except the last one I that may stop before, we have

$$\Delta(T) = \sum_{i=1}^{I(T)} \Delta_i\,, \tag{11.8}$$

where $I(T) = \llcorner \log_2(T+1) \lrcorner$. We further decompose $\Delta_i = \Delta_{1,i} + \Delta_{i,2}$ into the regret corresponding to the exploration stage $\Delta_{1,i}$ and the regret corresponding to the exploitation stage $\Delta_{i,2}$.

$\tau_{i,1}$ is the total length of the exploration stage i and $\tau_{i,2}$ is the total length of the exploitation stage i. For each model φ, we write $\tau_{i,1,J} \stackrel{\text{def}}{=} \frac{\tau_{i,1}}{J}$ the number of consecutive steps during which the UCRL2 algorithm is run with model φ in the exploration stage i, and $\tau_{i,2}(\varphi)$ the number of consecutive steps during which the UCRL2 algorithm is run with model φ in the exploitation stage i.

Good and Bad models. Let us now introduce the two following sets of models, defined after the end of the exploration stage, i.e. at time t_i.

$$\begin{aligned}\mathcal{G}_i &\stackrel{\text{def}}{=} \{\varphi \in \Phi \ ; \ \widehat{\mu}_{i,1}(\varphi) - 2B(i,\varphi,\delta) \geq \widehat{\mu}_{i,1}(\varphi^\star) - 2B(i,\varphi^\star,\delta)\} \backslash \{\varphi^*\}\,,\\ \mathcal{B}_i &\stackrel{\text{def}}{=} \{\varphi \in \Phi \ ; \ \widehat{\mu}_{i,1}(\varphi) - 2B(i,\varphi,\delta) < \widehat{\mu}_{i,1}(\varphi^\star) - 2B(i,\varphi^\star,\delta)\}\,.\end{aligned}$$

With this definition, we have the decomposition $\Phi = \mathcal{G}_i \cup \{\varphi^\star\} \cup \mathcal{B}_i$.

5.2.1 Regret in the exploration phase

Since in the exploration stage i each model φ is run for $\tau_{i,1,J}$ many steps, the regret for each model $\varphi \neq \varphi^\star$ is bounded by $\tau_{i,1,J}\rho^\star$. Now the regret for the true model is $\tau_{i,1,J}(\rho^\star - \widehat{\mu}_1(\varphi^\star))$, thus the total contribution to the regret in the exploration stage i is upper-bounded by

$$\Delta_{i,1} \leqslant \tau_{i,1,J}(\rho^\star - \widehat{\mu}_1(\varphi^\star)) + (J-1)\tau_{i,1,J}\rho^\star\,. \tag{11.9}$$

5.2.2 Regret in the exploitation phase

By definition, all models in $\mathcal{G}_i \cup \{\varphi^\star\}$ are selected before any model in $\mathcal{B}_i$ is selected.

The good models. Let us consider some $\varphi \in \mathcal{G}_i$ and an event Ω_i under which the exploitation phase does not reset. The test (equation (11.4)) starts after $\tau_{i,1,J}$, thus, since there is not reset, either $\tau_{i,2}(\varphi) = \tau_{i,1,J}$ in which case the contribution to the regret is bounded by $\tau_{i,1,J}\rho^\star$, or $\tau_{i,2}(\varphi) > \tau_{i,1,J}$, in which case the regret during the $(\tau_{i,2}(\varphi)-1)$ steps (where the test was successful) is bounded by

$$\begin{aligned}(\tau_{i,2}(\varphi)-1)(\rho^\star - \widehat{\mu}_{i,2,\tau_{i,2}(\varphi)-1}(\varphi)) &\leqslant (\tau_{i,2}(\varphi)-1)(\rho^\star - \widehat{\mu}_{i,1}(\varphi) + 2B(i,\varphi,\delta))\\ &\leqslant (\tau_{i,2}(\varphi)-1)(\rho^\star - \widehat{\mu}_{i,1}(\varphi^\star) + 2B(i,\varphi^\star,\delta))\,,\end{aligned}$$

and now since in the last step φ fails to pass the test, this adds a contribution to the regret at most $\rho^\star$.

We deduce that the total contribution to the regret of all the models $\varphi \in \mathcal{G}_i$ in the exploitation stages on the event Ω_i is bounded by

$$\Delta_{i,2}(\mathcal{G}_i) \leqslant \sum_{\varphi\in\mathcal{G}} \max\{\tau_{i,1,J}\rho^\star, (\tau_{i,2}(\varphi)-1)(\rho^\star - \widehat{\mu}_{i,1}(\varphi^\star) + 2B(i,\varphi^\star,\delta)) + \rho^\star\}\,. \tag{11.10}$$

The true model. First, let us note that since the total regret of the true model during the exploitation step i is given by $\tau_{i,2}(\varphi^\star)(\rho^\star - \widehat{\mu}_{i,2,t}(\varphi^\star))$, then the total regret of the exploration and exploitation stages in episode i on Ω_i is bounded by

$$\begin{aligned}\Delta_i \ \leqslant\ & \tau_{i,1,J}(\rho^\star - \widehat{\mu}_1(\varphi^\star)) + \tau_{i,1,J}(J-1)\rho^\star + \tau_{i,2}(\varphi^\star)(\rho^\star - \widehat{\mu}_{i,2,t_i+\tau_{i,2}}(\varphi^\star)) +\\ & \sum_{\varphi\in\mathcal{G}_i} \max\{\tau_{i,1,J}\rho^\star, (\tau_{i,2}(\varphi)-1)(\rho^\star - \widehat{\mu}_{i,1}(\varphi^\star) + 2B(i,\varphi^\star,\delta)) + \rho^\star\} + \sum_{\varphi\in\mathcal{B}_i} \tau_{i,2}(\varphi)\rho^\star\,.\end{aligned}$$

Now from the analysis provided in Jaksch et al. (2010) we know that when we run the UCRL2 with the true model $\varphi^\star$ with parameter $\delta_i(\delta)$, then there exists an event $\Omega_{1,i}$ of probability at least $1-\delta_i(\delta)$ such that on this event

$$\rho^\star - \widehat{\mu}_{i,1}(\varphi^\star) \leqslant B_D(\tau_{i,1,J}, \varphi^\star, \delta_i(\delta))\,,$$

and similarly there exists an event $\Omega_{2,i}$ of probability at least $1-\delta_i(\delta)$, such that on this event

$$\rho^\star - \widehat{\mu}_{i,2,t}(\varphi^\star) \leqslant B_D(\tau_{i,2}(\varphi^\star), \varphi^\star, \delta_1(\delta))\,.$$

Now we show that, with high probability, the true model $\varphi^\star$ passes all the tests (equation (11.4)) until the end of the episode i, and thus equivalently, with high probability no model $\varphi \in \mathcal{B}_i$ is selected, so that $\sum_{\varphi\in\mathcal{B}_i} \tau_{i,2}(\varphi) = 0$.

For the true model, after $\tau(\varphi^\star, t) \geqslant \tau_{i,1,J}$, there remains at most $(\tau_{i,2} - \tau_{i,1,J} + 1)$ possible timesteps where we do the test for the true model $\varphi^\star$. For each test we need to control $\mu_{i,2,t}(\varphi^\star)$, and the event corresponding to $\widehat{\mu}_{i,1}(\varphi^\star)$ is shared by all the tests. Thus we deduce that with probability higher than $1-(\tau_{i,2}-\tau_{i,1,J}+2)\delta_i(\delta)$ we have simultaneously on all time step until the end of exploitation phase of stage i,

$$\begin{aligned}
\widehat{\mu}_{i,2,t}(\varphi^\star) - \widehat{\mu}_{i,1}(\varphi^\star) &= \widehat{\mu}_{i,2,t}(\varphi^\star) - \rho^\star + \rho^\star - \widehat{\mu}_{i,1}(\varphi^\star) \\
&\geqslant -B_D(\tau(\varphi^\star,t), \varphi^\star, \delta_i(\delta)) - B_D(\tau_{i,1,J}, \varphi^\star, \delta_i(\delta)) \\
&\geqslant -2B_D(\tau_{i,1,J}, \varphi^\star, \delta_i(\delta))\,.
\end{aligned}$$

Now provided that $f(t_i) \geqslant D$, then $B_D(\tau_{i,1,J}, \varphi^\star, \delta_i(\delta)) \leqslant B(i, \varphi^\star, \delta)$, thus the true model passes all tests until the end of the exploitation part of stage i on an event $\Omega_{3,i}$ of probability higher than $1-(\tau_{i,2}-\tau_{i,1,J}+2)\delta_i(\delta)$. Since there is no reset, we can choose $\Omega_i \overset{\text{def}}{=} \Omega_{3,i}$. Note that on this event, we thus have $\sum_{\varphi\in\mathcal{B}_i} \tau_{i,2}(\varphi) = 0$.

By using a union bound over the events $\Omega_{1,i}, \Omega_{2,i}$ and $\Omega_{3,i}$, then we deduce that with probability higher than $1-(\tau_{i,2}-\tau_{i,1,J}+4)\delta_i(\delta)$,

$$\begin{aligned}
\Delta_i \;\leqslant\; & \tau_{i,1,J} B_D(\tau_{i,1,J}, \varphi^\star, \delta_i(\delta))) + [\tau_{i,1,J}(J-1) + |\mathcal{G}_i|]\rho^\star + \tau_{i,2}(\varphi^\star) B_D(\tau_{i,2}(\varphi^\star), \varphi^\star, \delta_i(\delta)) \\
& + \sum_{\varphi\in\mathcal{G}_i} \max\{(\tau_{i,1,J}-1)\rho^\star, (\tau_{i,2}(\varphi)-1)(B_D(\tau_{i,1,J}, \varphi^\star, \delta_i(\delta)) + 2B(i,\varphi^\star,\delta)\}\,.
\end{aligned}$$

Now using again the fact that $f(t_i) \geqslant D$, and after some simplifications, we deduce that

$$\begin{aligned}
\Delta_i \;\leqslant\; & \tau_{i,1,J} B_D(\tau_{i,1,J}, \varphi^\star, \delta_i(\delta)) + \tau_{i,2}(\varphi^\star) B_D(\tau_{i,2}(\varphi^\star), \varphi^\star, \delta_i(\delta)) \\
& + \sum_{\varphi\in\mathcal{G}_i} (\tau_{i,2}(\varphi)-1) 3B(i,\varphi^\star,\delta) + \tau_{i,1,J}(J + |\mathcal{G}_i| - 1)\rho^\star\,.
\end{aligned}$$

Finally, we use the fact that $\tau B_D(\tau, \varphi^\star, \delta_i(\delta))$ is increasing with τ to deduce the following rough bound that holds with probability higher than $1-(\tau_{i,2}-\tau_{i,1,J}+4)\delta_i(\delta)$

$$\Delta_i \;\leqslant\; \tau_{i,2} B(i,\varphi^\star,\delta) + \tau_{i,2} B_D(\tau_{i,2}, \varphi^\star, \delta_i(\delta)) + 2J\tau_{i,1,J}\rho^\star\,,$$

where we used the fact that $\tau_{i,2} = \tau_{i,2}(\varphi^\star) + \sum_{\varphi\in\mathcal{G}} \tau_{i,2}(\varphi)\,.$

5.3 Tuning the parameters of each stage.

We now conclude by tuning the parameters of each stage, i.e. the probabilities $\delta_i(\delta)$ and the length τ_i, $\tau_{i,1}$ and $\tau_{i,2}$. The total length of stage i is by definition

$$\tau_i = \tau_{i,1} + \tau_{i,2} = \tau_{i,1,J} J + \tau_{i,2}\,,$$

where $\tau_i = 2^i$. So we set $\tau_{i,1} \stackrel{\text{def}}{=} \tau_i^{2/3}$ and then we have $\tau_{i,2} \stackrel{\text{def}}{=} \tau_i - \tau_i^{2/3}$ and $\tau_{i,1,J} = \frac{\tau_i^{2/3}}{J}$. Now using these values and the definition of the bound $B(i,\varphi^\star,\delta)$, and $B_D(\tau_{i,2},\varphi^\star,\delta_i(\delta))$, we deduce with probability higher than $1-(\tau_{i,2}-\tau_{i,1,J}+4)\delta_i(\delta)$ the following upper bound

$$\Delta_i \;\leqslant\; 34 f(t_i) S \sqrt{AJ \log\Big(\frac{\tau_i^{2/3}}{J\delta_i(\delta)}\Big)\tau_i^{2/3}} + 34 DS \sqrt{A \log\Big(\frac{\tau_i}{\delta_i(\delta)}\Big)\tau_i} + 2\tau_i^{2/3}\rho^\star\,,$$

with $t_i = 2^i - 1 + 2^{2i/3}$ and where we used the fact that $\Big(\frac{J}{\tau_i^{2/3}}\Big)^{1/2}\tau_{i,2} \leqslant \sqrt{J}\tau_i^{2/3}$.

We now define $\delta_i(\delta)$ such that $\delta_i(\delta) \stackrel{\text{def}}{=} (2^i - (J^{-1}+1)2^{2i/3}+4)^{-1} 2^{-i+1}\delta$.

Since for the stages $i \in \mathcal{I}_0 \stackrel{\text{def}}{=} \{i \geqslant 1; f(t_i) < D\}$, the regret is bounded by $\Delta_i \leqslant \tau_i \rho^\star$, then the total cumulative regret of the algorithm is bounded with probability higher than $1-\delta$ (using the defition of the $\delta_i(\delta)$) by

$$\Delta(T) \leqslant \sum_{i \notin \mathcal{I}_0} [34 f(t_i) S \sqrt{JA \log\Big(\frac{2^{8i/3}}{J\delta}\Big)} + 2] 2^{2i/3} + 34 DS \sqrt{A \log\Big(\frac{2^{3i}}{\delta}\Big) 2^i} + \sum_{i \in \mathcal{I}_0} 2^i \rho^\star\,.$$

where $t_i = 2^i - 1 + 2^{2i/3} \leqslant T$.

We conclude by using the fact that since $I(T) \leqslant \log_2(T+1)$, then with probability higher than $1-\delta$, the following bound on the regret holds

$$\Delta(T) \leqslant c f(T) S \Big(AJ \log(J\delta)^{-1} \log_2(T)\Big)^{1/2} T^{2/3} + c' DS \Big(A \log(\delta^{-1}) \log_2(T) T\Big)^{1/2} + c(f,D)\,.$$

for some constant c, c', and where $c(f,D) = \sum_{i \in \mathcal{I}_0} 2^i \rho^\star$. Now for the special choice when $f(T) \stackrel{\text{def}}{=} \log_2(T+1)$, then $i \in \mathcal{I}_0$ means $2^i + 2^{2i/3} < 2^D + 2$, thus we must have $i < D$, and thus $c(f,d) \leqslant 2^D$.

Chapter 12

Perspectives and Future Work.

In this concluding section, I would like to briefly mention some of the immediate future works already planned after this PhD thesis. First, there are obviously some natural extensions to the chapters presented here.

- Following chapter 2 and the work by Garivier and Cappé (2011), it is natural to extend the use of the Kullback-Leibler divergence in bandit theory to other classes of distributions. This would require techniques coming from transport theory and non trivial extensions of Sanov's Theorem.

- It seems natural to extend the ideas developped in chapter 4 and chapter 11 about model selection or model aggregation in bandit, game theory and reinforcement learning. This will require developing tools for non-asymptotic hypothesis testing, and is also linked with code theory as well as random graphs, and definitely opens a large avenue of research about the question of *adaptivity* in sequential learning.

- Also it seems natural to apply recent advances from PAC analysis to bandits, and continue developing this field of research, see chapter 5.

- Chapter 6 studies the use of Gaussian random matrices for regression, that are designed in a non-adaptive way. It seems natural to extend this idea to data-driven matrix generation, which would need a more intricate analysis due to the fact that this introduces a dependency. Of course the same idea, i.e. to use a data-dependent random operator, applies to chapter 7 as well.

- It would be very nice to apply results for regression with Markov design such as those developped by Stéphane Gaiffas to the TD algorithm, as it would enable to both generalize and unify chapter 9 and chapter 10.

Of course there are plenty of other perspectives of research, including studying the effect of adding a numerical cost to the notion of regret in bandit, addressing the cover-shift problem in reinforcement learning, working on inverse reinforcement learning as well as topics raised in the foreword chapter...etc. I would really like to work on the scarcely addressed problem of cooperative bandits, from a non-asymptotic and adaptive perspective, all the more so that there is a big practical motivation coming from Brain Machine Interfaces as well as few theoretical results.

Finally, better understanding the statistical properties of the empirical processes involved in regression or in reinforcement learning by means of local central limit theorems and adaptive confidence bounds (for instance) is a deep, interesting question that would allow for much more precise statements than only current first-order analysis. This is left for future work.

More generally, developing the right theoretical tools that enable to both address such questions and at the same time can be used in practical algorithms is actually challenging and needs communication between seemingly distant areas of research, which is difficult. However I believe that we really need strong ideas both practical and theoretical in order to succeed in such a task, and thus strong communication as well.

Summary of Scientific Activity.

Here is a brief overview of the research activity during these three years.

Publications. Note that conference papers and journals currently under review do not appear in this list.

2011

- Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences, Maillard et al. (2011).
- Adaptive bandits: Towards the best history-dependent strategy, Maillard and Munos (2011).

2010

- Finite-Sample Analysis of Bellman Residual Minimization, Maillard et al. (2010).
- Scrambled Objects for Least-Squares Regression, Maillard and Munos (2010a).
- LSTD with Random Projections, Ghavamzadeh et al. (2010a).
- Online Learning in Adversarial Lipschitz Environments, Maillard and Munos (2010b).

2009

- Compressed Least Squares Regression, Maillard and Munos (2009).
- Complexity versus Agreement for Many Views, Maillard and Vayatis (2009).

Broadcast. Finally, as a service to the profession, I created the google group "Probability and Statistics news" [PS-news:`http://groups.google.fr/group/maths-ps-news`] in order to help broadcasting job announcements or conference events related to mathematical probability and statistics, like the google group "Machine-Learning news" [ML-news:`http://groups.google.fr/group/ml-news`] does successfully for the machine learning community. The goal is here to provide a tool in order to facilitate inter and intra-communication for the two strong communities of Probability and of Statistics at a world-scale level.

Bibliography

Boularias Abdeslam, Kober Jens, and Peters Jan. Relative entropy inverse reinforcement learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the 14th international conference on Artificial Intelligence and Statistics*, AI&Stats '11, JMLR W&CP, 2011. xx

Jacob D. Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In Servedio and Zhang (2008). 22, 65

Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Servedio and Zhang (2008), pages 263–274. xxx, 22, 63, 64

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, June 2003. 115, 136

René Aïd, Vincent Grellier, Arnaud Renaud, and Olivier Teytaud. Application de l'apprentissage par renforcement à la gestion du risque. In *Conférences Francophone sur l'Apprentissage Automatique*, 2003. xvi

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the 38th annual ACM Symposium on Theory of computing*, STOC '06, pages 557–563, New York, NY, USA, 2006. ACM. ISBN 1-59593-134-1. 128

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In José L. Balcázar, Philip M. Long, and Frank Stephan, editors, *Proceedings of the 17th international conference on Algorithmic Learning Theory*, volume 4264 of *ALT '06, Lecture Notes in Computer Science*, pages 229–243, Barcelona, Spain, oct 2006. Springer. ISBN 3-540-46649-5. 20

Pierre Alquier. PAC-Bayesian Bounds for Randomized Empirical Risk Minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, dec 2008. 118, 140

Pierre Alquier and Karim Lounici. PAC-Bayesian Bounds for Sparse Regression Estimation with Exponential Weights. Technical Report 2010-40, Centre de Recherche en Economie et Statistique, 2010. To appear. 141

Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical monographs*. Oxford University Press, 2000. 111

Rie Kubota Ando and Tong Zhang. Learning on graph with laplacian regularization. In Platt et al. (2007), pages 25–32. 182

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michaël I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50:5–43, jan 2003. 69

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008. 208, 213, 218, 236, 238

Jean-Marie Aubry and Stéphane Jaffard. Random wavelet series. *Communications in Mathematical Physics*, 227:483–514, 2002. 130

Jean-Yves Audibert and Olivier Bousquet. Combining pac-bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8:863–889, December 2007. 116, 121

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In Dasgupta and Klivans (2009). 85

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, December 2010. xxix, 11, 12, 37, 40

Jean-Yves Audibert and Olivier Catoni. Robust linear regression through PAC-Bayesian truncation. 2010a. 140, 142

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. 48 pages 62J05, 62J07, 2010b. 117, 118, 119

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410: 1876–1902, 2009. xxix, 11, 12, 37, 40, 85

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, March 2003. ISSN 1532-4435. 8, 22, 64

Peter Auer and Ron Meir, editors. volume 3559 of *COLT '05, Lecture Notes in Computer Science*, Bertinoro, Italy, jun 2005. Springer. ISBN 3-540-26556-2. 261, 266

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *Proceedings of the 20th conference on advances in Neural Information Processing Systems*, NIPS '06, pages 49–56, Vancouver, British Columbia, Canada, dec 2006. MIT Press. ISBN 0-262-19568-2. xxi

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010. xxix, 37, 40

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th annual symposium on Foundations of Computer Science*, FOCS '95, pages 322–331, Milwaukee, WI, USA, 1995. IEEE Computer Society Press. ISBN 0-8186-7183-1. 17, 19, 64, 117, 248

Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:2002, 2000. 68

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, may 2002. ISSN 0885-6125. xxix, 10, 37, 40, 81, 84, 243

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, January 2003. ISSN 0097-5397. 24, 29, 64, 74, 81, 85, 88, 91

Baruch Awerbuch and Robert D. Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer Systems and Science*, 74:97–114, February 2008. ISSN 0022-0000. 21, 22

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367, 1967. 112

Leemon C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, ICML '95, pages 30–37, 1995. 208

Richard G. Baraniuk, Mark A. Davenport, Ronald DeVore, and Michaël B. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008. 167, 168, 173

Andrew Barron, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Approximation and learning by greedy algorithms. 36:1:64–94, 2008. 128, 147

Peter L. Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the 25th conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 35–42, Arlington, Virginia, United States, 2009. AUAI Press. ISBN 978-0-9749039-5-8. 242, 244, 249

Peter L. Bartlett, Michaël I. Jordan, and Jon D. Mcauliffe. Convexity, classification, and risk bounds. Technical report, Journal of the American Statistical Association, 2003. 183

Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. In Platt et al. (2007), pages 65–72. 22, 29, 65

Peter L. Bartlett, Varsha Dani, Thomas P. Hayes, Sham M. Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In Servedio and Zhang (2008), pages 335–342. 21

Jonathan Baxter, Andrew Tridgell, and Lex Weaver. *Reinforcement learning and chess*, pages 91–116. Nova Science Publishers, Inc., Commack, NY, USA, 2001. xvi

Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In John Shawe-Taylor and Yoram Singer, editors, *Proceedings of the 17th annual Conference On Learning Theory*, volume 3120 of *COLT '04, Lecture Notes in Computer Science*, pages 624–638, Banff, Canada, jul 2004. Springer. ISBN 3-540-22282-0. 182

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. On Manifold Regularization. In *Proceedings of the 8th international conference on Artificial Intelligence and Statistics*, AI&Stats '05, 2005. 181, 182, 184, 193

Shai Ben-David and Ulrike von Luxburg. Relating clustering stability to properties of cluster boundaries. In Servedio and Zhang (2008), pages 379–390. 199

Shai Ben-David, Ulrike von Luxburg, and David Pál. A sober look at clustering stability. In Lugosi and Simon (2006), pages 5–19. ISBN 3-540-35294-5. 199

Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Chris K. I. Williams, and Aron Culotta, editors. NIPS '09, Vancouver, British Columbia, Canada, dec 2009. 271

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. 174

Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *Annals of Applied Probability*, 18(5):1848–1869, 2008. 113

Daniel S. Bernstein and Shlomo Zilberstein. Reinforcement learning for weakly coupled mdps and an application to planetary rover control. 2001. xvi

Sergei Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Original publication: Ann. Sci. Inst. Sav. Ukraine, Sect. Math. 1*, 3(1), 1924. 108

Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp. Bandit problems with infinitely many arms. *Annals of Statistics*, (25):2103–2116, 1997. 21

Philippe Berthet and Zhan Shi. Small ball estimates for brownian motion under a weighted sup-norm. In *Studia Sci. Math. Hung*, volume 36, pages 1–2, 275–289, 2001. 199

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control (The Discrete Time Case)*. Academic Press, New York, 1978. 208

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996. 205, 208, 209

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, sep 1998. 141

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, aug 2006. ISBN 0387310738. 72, 193

Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003. 199

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36:489–531, Apr 2008. 183, 199

Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning*, ICML '01, pages 19–26, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. 180

Avrim Blum and Yishay Mansour. From external to internal regret. In Auer and Meir (2005), pages 621–636. ISBN 3-540-26556-2. 24, 81, 82, 83

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, December 2007. 81

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th annual conference on Computational Learning theory*, COLT '98, pages 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0. 180, 181

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: P&S*, 9:323–375, 2005. 186

Gerard Bourdaud. Ondelettes et espaces de besov. *Rev. Mat. Iberoamericana*, 11:3:477–512, 1995. 134

Justin A. Boyan. Least-squares temporal difference learning. *Proceedings of the 16th International Conference on Machine Learning*, pages 49–56, 1999. xxi, 226

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning Journal*, 22:33–57, 1996. 226

Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, March 2003. ISSN 1532-4435. 242

Ulf Brefeld, Thomas Gärtner, Tobias Scheffer, and Stefan Wrobel. Efficient co-regularised least squares regression. In Cohen and Moore (2006), pages 137–144. ISBN 1-59593-383-2. 181

Sébastien Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université de Lille 1, 2010. 85, 86, 93, 95, 96, 98, 99

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization of X-armed bandits. In Koller et al. (2008). 22, 65

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In Gavaldà et al. (2009), pages 23–37. ISBN 978-3-642-04413-7. 24

Hans-Joachim Bungartz and Michaël Griebel. Sparse grids. In Arieh Iserles, editor, *Acta Numerica*, volume 13. University of Cambridge, 2004. 150

Apostolos N. Burnetas and Michaël N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996. xxix, xxx, 4, 8, 9, 10, 13, 14, 37, 39, 40, 43, 51

E.J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008. 164

Emmanuel J. Candés and Justin K. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23:969–985, 2007. 163, 164

Emmanuel J. Candés and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007. 164

Emmanuel J. Candés, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006a. 164

Emmanuel J. Candés, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006b. 164

Emmanuel J. Candés, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *CoRR*, abs/0912.3599, 2009. xix

Stéphane Canu, Xavier Mary, and Alain Rakotomamonjy. Functional learning through kernel. *arXiv*, oct 2009. 133, 135

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, jul 2007. ISSN 1615-3375. 128, 141

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization*. Springer-Verlag, 2004. 140

Alain Celisse. *Model selection via cross-validation in density estimation, regression, and change-points detection*. PhD thesis, Université Paris Sud, Faculté des Sciences d'Orsay, December 2008. 199

Nicolò Cesa-Bianchi and Gábor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Journal of Machine Learning*, 51(3):239–261, 2003. ISSN 0885-6125. 24, 81

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089. 64, 66, 68

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. In Dasgupta and Klivans (2009). xxx, 22, 63, 64

Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. In *Proceedings of the 25th annual ACM Symposium on Theory Of Computing*, STOC '93, pages 382–391, New York, NY, USA, 1993. ACM. ISBN 0-89791-591-7. 19

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44:427–485, May 1997. ISSN 0004-5411. 66

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51:77–92, 2005. 20

Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. In Koller et al. (2008), pages 273–280. xvii, 23

Doran Chakraborty and Peter Stone. Online model learning in adversarial markov decision processes. In *Proceedings of the 9th international conference on Autonomous Agents and Multiagent Systems: volume 1 - Volume 1*, AAMAS '10, pages 1583–1584, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-0-9826571-1-9. xxi

Olivie Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006. 180

Yuan S. Chow and Henry Teicher. *Probability Theory*. Springer, 1988. 11, 45

Chih chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50:338–355, 2005. xvii, 19

William W. Cohen and Andrew Moore, editors. volume 148 of *ICML '06, ACM International Conference Proceeding Series*, Pittsburgh, Pennsylvania, USA, jun 2006. ACM. ISBN 1-59593-383-2. 262, 269

William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors. volume 307 of *ICML '08, ACM International Conference Proceeding Series*, Helsinki, Finland, jun 2008. ACM. ISBN 978-1-60558-205-4. 269, 275

Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the 19th annual ACM Symposium on Theory 0f computing*, STOC '87, pages 1–6, New York, NY, USA, 1987. ACM. 150

Rémi Coulom. Computing Elo ratings of move patterns in the game of Go. *ICGA Journal*, 30(4):198–208, December 2007. xvi

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991. 111

Harald Cramér. Sur un nouveau théorème limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles*, 736:5–23, 1938. 115

Robert Crites and Andrew G. Barto. Improving elevator performance using reinforcement learning. In *Advances in Neural Information Processing Systems 8*, pages 1017–1023. MIT Press, 1996. xvi

Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12:768–793, 1984. 109, 110

A. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *Arxiv preprint arXiv:0903.1223*, 2009. 164

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning Journal*, 72:39–61, August 2008. 121

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. The price of bandit information for online optimization. In Koller et al. (2008), pages 345–352. xxx, 19, 22, 63, 64

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In Servedio and Zhang (2008), pages 355–366. 22, 64

Sanjot Dasgupta and Adam Klivans, editors. COLT '09, Montreal,Quebec, Canada, jun 2009. 258, 263, 270

Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the 40th annual ACM Symposium on Theory Of Computing*, STOC '08, pages 537–546, New York, NY, USA, 2008. ACM. 126

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22:60–65, jan 2003. 114

Victor H. De la Peña. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):pp. 537–564, jan 1999. 114

Victor H. de la Peña, Michaël J. Klass, and Tze Leung Lai. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902–1933, jul 2004. 114

Sébastien Deguy and Albert Benassi. A flexible noise model for designing maps. In *Proceedings of the Vision Modeling and Visualization Conference 2001*, VMV '01, pages 299–308. Aka GmbH, 2001. ISBN 3-89838-028-9. 130

Pierre Del Moral. *Feynman-Kac formulae : genealogical and interacting particle systems with applications*. New York : Springer, 2004. ISBN 0387202684. 70

Sylvain Delattre and Stéphane Gaiffas. Nonparametric regression with martingale increment errors. 2010. 114

Amir Dembo and Ofer Zeitouni. Large deviations techniques and applications. *Elearn*, 1998. 109, 110

Ronald DeVore. *Nonlinear Approximation*. Acta Numerica, 1997. 148

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996. 78

Ian H. Dinwoodie. Mesures dominantes et théorème de Sanov. *Annales de l'Institut Henri Poincaré – Probabilités et Statistiques*, 28(3):365–373, 1992. 41, 58, 110

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006. 164, 173, 174

David L. Donoho and Philip B. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989. 164

Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Minimum variance importance sampling via population monte carlo. *Esaim P&S*, 11, 2007. 70

Richard M. Dudley. *Real Analysis and Probability*. Wadsworth, Belmont, Calif, 1989. 124

Arnaud Durand. Random wavelet series based on a tree-indexed markov chain. *Communications in Mathematical Physics*, 283:451–477, 2008. 130

Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal of Matrix Analysis and Applications*, 9:543–560, December 1988. 115

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv; a reinforcement learning approach. In *Machine Learning Conference of Belgium and the Netherlands (Benelearn)*, pages 65–72, 2006. xvii, 80

Amir M. Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration. In Koller et al. (2008), pages 441–448. 208, 226, 235

Amir M. Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration for planning in continuous-space Markovian decision problems. In *Proceedings of the American Control Conference*, 2009. 226

Amir M. Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In Lafferty et al. (2010). 218

Alexei A. Fedotov, Peter Harremoës, and Flemming Topsøe. Best Pinsker Bound equals Taylor Polynomial of Degree 49. *Computational Technologies*, 8:3–14, 2003. 111

Sarah Filippi. *Stratégies optimistes en apprentissage par renforcement*. PhD thesis, Télécom ParisTech, 2010. xxx, 18, 37, 40, 43

Abraham D. Flaxman, Adam Tauman Kalai, and Hugh Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th annual ACM-SIAM Symposium On Discrete Algorithms*, SODA '05, pages 385–394. SIAM, 2005. 22, 65

Maria florina Balcan and Avrim Blum. A pac-style model for learning from labeled and unlabeled data. In Auer and Meir (2005), pages 111–126. ISBN 3-540-26556-2. 180

Massimo Fornasier and Hölger Rauhut. *Compressive Sensing*. Springer, to appear. 173

Dean P. Foster and Rakesh Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1996. 23, 81

Dean P. Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7–35, 1999. 24, 81

Simon Foucart and Ming-Jun Lai. Sparsest solutions of underdetermined linear systems via lq-minimization for $0 < q < p$. *Applied and Computational Harmonic Analysis*, 26(3): 395–407, 2009. 163, 167, 175, 177

Michaël Frazier and Björn Jawerth. Decomposition of besov spaces. *Indiana University Mathematics Journal*, (34), 1985. 134

David A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, 3: 100–118, 1975. 113

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT '95: Proceedings of the 2nd European conference on COmputational Learning Theory*, pages 23–37, London, UK, 1995. Springer-Verlag. ISBN 3-540-59119-2. 81

Johannes Fürnkranz and Thorsten Joachims, editors. ICML '10, Haifa, Israel, jun 2010. Omnipress. 270, 273, 274

Aurélien Garivier. Deviation bounds. Private communication, 2011. 113

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011. 37, 41, 56, 253

Aurélien Garivier and Florencia Leonardi. Context tree selection: A unifying view. arXiv:1011.2424, 2010. 42, 56, 112, 113

Ricard Gavaldà, Gábor Lugosi, Thomas Zeugmann, and Sandra Zilles, editors. volume 5809 of *ALT '09, Lecture Notes in Computer Science*, Porto, Portugal, oct 2009. Springer. ISBN 978-3-642-04413-7. 262, 271, 272

Sylvain Gelly and David Silver. Combining online and offline knowledge in uct. In Ghahramani (2007), pages 273–280. ISBN 978-1-59593-793-3. xvi

Zoubin Ghahramani, editor. volume 227 of *ICML '07, ACM International Conference Proceeding Series*, Corvalis, Oregon, USA, jun 2007. ACM. ISBN 978-1-59593-793-3. 267, 272

Mohammad Ghavamzadeh, Alessandro Lazaric, Odalric-Ambrym Maillard, and Rémi Munos. Lstd with random projections. In Lafferty et al. (2010), pages 721–729. 225, 255

Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, and Odalric-Ambrym Maillard. LSPI with random projections. Technical report, INRIA, 2010b. 234, 236

Wally R. Gilks, Sylvia Richardson, and David J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman Hall/CRC, Boca Raton, FL, 1996. 69

John C. Gittins, Richard Weber, and Kevin Glazebrook. *Multi-armed Bandit Allocation Indices*. Wiley, 1989. 25

Carl Gold, Alex Holub, and Peter Sollich. Bayesian approach to feature selection and parameter tuning for support vector machine classifiers. *Neural Network*, 18(5-6):693–701, 2005. ISSN 0893-6080. 199

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, oct 1996. ISBN 0801854148. 190

Nathan Gozlan and Christian Léonard. Transport inequalities - a survey. *Markov Processes and Related Fields*, pages 635–736, 2010. 111

Sudipto Guha and Kamesh Munagala. Approximation algorithms for budgeted learning problems. In *Proceedings of ACM Symposium on Theory of Computing*, pages 104–113, 2007. 21

Sudipto Guha, Kamesh Munagala, and Saswati Sarkar. Information acquisition and exploitation in multichannel wireless systems. *IEEE Transactions on Information Theory*, 2007a. 18

Sudipto Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. *CoRR*, abs/0711.3861, 2007b. 18

László Györfi, Michaël Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002. 140, 143, 145, 146, 154, 212, 215

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000. ISSN 00129682. 24, 81

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. In Lugosi and Simon (2006), pages 499–513. ISBN 3-540-35294-5. xxx, 22, 63, 64

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. 107, 108

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In Adam Tauman Kalai and Mehryar Mohri, editors, *Proceedings of the 23rd annual Conference On Learning Theory*, pages 67–79. Omnipress, June 2010a. ISBN 978-0-9822529-2-5. xxx, 10, 37, 40, 51, 53

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. arXiv:0905.2776, 2010b. 40, 49, 50, 59, 60, 61

Marcus Hutter. Feature reinforcement learning: Part I: Unstructured MDPs. *Journal of Artificial General Intelligence*, 1:3–24, oct 2009. 82, 248

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010. ISSN 1532-4435. 205, 242, 243, 244, 248, 249, 251

Svante Janson. *Gaussian Hilbert spaces*. Cambridge Univerity Press, Cambridge, UK, 1997. 131, 136

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In Cohen et al. (2008), pages 440–447. ISBN 978-1-60558-205-4. xxx, 22, 63, 64

Satyen Kale, Lev Reyzin, and Robert E. Schapire. Non-stochastic bandit slate problems. In Lafferty et al. (2010), pages 1054–1062. 21

Varun Kanade, H. Brendan McMahan, and Brent Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Proceedings of the 12th international conference on Artificial Intelligence and Statistics*, number 5 in AI&Stats '09, pages 272–279, 2009. 23, 82, 87

Michaël Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning Journal*, 49:209–232, November 2002. ISSN 0885-6125. 242

Philipp W. Keller, Shie Mannor, and Doina Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In Cohen and Moore (2006), pages 449–456. ISBN 1-59593-383-2. 226

Robert D. Kleinberg, Alexandru Niculescu-mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In Servedio and Zhang (2008), pages 425–436. 22, 23, 65, 87

Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors. NIPS '08, Vancouver, British Columbia, Canada, dec 2008. MIT Press. 262, 263, 264, 266, 277

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 141, 199, 200

Vladimir Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15 (3):799–828, 2009. 164

Zico Kolter and Andrew Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, volume 382 of *ICML '09, ACM International Conference Proceeding Series*, pages 521–528, Montreal, Quebec, Canada, jun 2009. ACM. ISBN 978-1-60558-516-1. 226

Risi I. Kondor and John D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*, ICML '02, pages 315–322, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1-55860-873-7. 182

John D. Lafferty, Chris K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors. NIPS '10, Vancouver, British Columbia, Canada, dec 2010. 266, 267, 269, 271, 272

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003. 208, 218, 226

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985. xxix, 4, 8, 9, 11, 13, 37, 39, 40, 243

Alessandro Lazaric and Rémi Munos. Hybrid stochastic-adversarial online learning. In Dasgupta and Klivans (2009). 19

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of LSTD. Technical Report inria-00482189, INRIA, 2010a. xxi, 105, 228, 229, 232

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. Technical Report inria-00528596, INRIA, 2010b. 230, 232, 234, 236, 237, 238

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of LSTD. In Fürnkranz and Joachims (2010). 208, 219

Michel Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001. 107

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds Mach.*, 17:391–444, December 2007. xiii

Ehud Lehrer and Dinah Rosenberg. A wide range no-regret theorem. Game theory and information, EconWPA, 2003. 81

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008. 69

Wenbo V. Li and Werner Linde. Approximation, metric entropy and small ball estimates for gaussian measures. *Annals of Probability*, 27:1556–1578, 1999. 199

Edo Liberty, Nir Ailon, and Amit Singer. Dense fast random projections and lean walsh transforms. In *Proceedings of the 11th international workshop, APPROX 2008, and 12th*

international workshop, RANDOM 2008 on Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques, APPROX '08 / RANDOM '08, pages 512–522, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85362-6. 128

Mikhail A. Lifshits. *Gaussian random functions*. Kluwer Academic Publishers, Dordrecht, Boston, 1995. 131, 132

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. In *Proceedings of the 30th annual Symposium on Foundations of Computer Science*, pages 256–261, Washington, DC, USA, 1989. IEEE Computer Society. ISBN 0-8186-1982-1. 19, 66

Manuel Loth, Manuel Davy, and Philippe Preux. Sparse temporal difference learning using lasso. In *IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 352–359, 2007. 226

Tyler Lu, David Pál, and Martin Pál. Contextual multi-armed bandits. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the 13th international conference on Artificial Intelligence and Statistics*, volume 9, pages 485–492, 2010. xvii, 19

Gábor Lugosi and Hans-Ulrich Simon, editors. volume 4005 of *COLT '06, Lecture Notes in Computer Science*, Pittsburgh, PA, USA, jun 2006. Springer. ISBN 3-540-35294-5. 260, 268

Sridhar Mahadevan. Representation policy iteration. In *Proceedings of the 21st conference on Uncertainty in Artificial Intelligence*, UAI '05, pages 372–379, 2005. 226

Odalric-Ambrym Maillard and Rémi Munos. Compressed least-squares regression. In Bengio et al. (2009), pages 1213–1221. 123, 125, 137, 149, 230, 231, 255

Odalric-Ambrym Maillard and Rémi Munos. Scrambled objects for least-squares regression. In Lafferty et al. (2010), pages 1549–1557. 123, 125, 235, 255

Odalric-Ambrym Maillard and Rémi Munos. Online learning in adversarial lipschitz environments. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD'10, pages 305–320, Berlin, Heidelberg, 2010b. Springer-Verlag. ISBN 3-642-15882-X, 978-3-642-15882-7. 22, 29, 63, 255

Odalric-Ambrym Maillard and Rémi Munos. Adaptive bandits: Towards the best history-dependent strategy. In *To appear in Proceedings of the 14th international conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR W&CP*, 2011. 20, 24, 79, 255

Odalric-Ambrym Maillard and Nicolas Vayatis. Complexity versus agreement for many views. In Gavaldà et al. (2009), pages 232–246. ISBN 978-3-642-04413-7. 179, 255

Odalric-Ambrym Maillard, Rémi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh. Finite sample analysis of bellman residual minimization. In *Asian Conference on Machine Learning*, 2010. 207, 255

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *To appear in Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011. 37, 255

Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, apr 1967. 115

Ishai Menache, Shie Mannor, and Nahum Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 134:215–238, 2005. 226

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 19th International Conference on Machine Learning*, ICML '03, pages 560–567, 2003. 208, 218, 220

Rémi Munos. Performance bounds in Lp norm for approximate value iteration. *SIAM Journal of Control and Optimization*, 2007. 208

Rémi Munos and Csaba Szepesvári. Finite time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008. 218, 238

Hariharan Narayanan. Randomized interior point methods for sampling and optimization. *Computing Research Repository*, 2009. 70

Hariharan Narayanan and Alexander Rakhlin. Random walk approach to regret minimization. In Lafferty et al. (2010), pages 1777–1785. xxx, 22, 29, 31, 70, 116

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. xx

Ronald Ortner. Online regret bounds for markov decision processes with deterministic transitions. In Gavaldà et al. (2009), pages 123–137. ISBN 978-3-642-04413-7. 83

Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal. Multi-armed bandit problems with dependent arms. In Ghahramani (2007). ISBN 978-1-59593-793-3. 21

Ronald Parr, Christopher Painter-wakefield, Lihong Li, and Michaël L. Littman. Analyzing feature generation for value-function approximation. In Ghahramani (2007), pages 737–744. ISBN 978-1-59593-793-3. 226

Dmitry Pechyony, Rauf Izmailov, Akshay Vashist, and Vladimir N. Vapnik. Smo-style algorithms for learning using privileged information. In *DMIN*, pages 235–241, 2010. 18

Marek Petrik, Gavin Taylor, Ronald Parr, and Shlomo Zilberstein. Feature selection using regularization in approximate linear programs for Markov decision processes. In Fürnkranz and Joachims (2010), pages 871–878. 226

John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors. NIPS '07, Vancouver, British Columbia, Canada, dec 2007. MIT Press. 258, 260, 273, 276

Jan Poland. Nonstochastic bandits: Countable decision set, unbounded costs and reactive environments. *Theoretical Computuer Science*, 397(1-3):77–93, jul 2008. 21, 64

David Pollard. *Empirical processes: theory and applications*. NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, 1990. 109

Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2007. 208, 218

Martin L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994. 208

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In Platt et al. (2007). 148

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. 2008. 148, 149

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Beyond regret. *ArXiv e-prints*, nov 2010. 25, 105, 179

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2586–2591, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. xx

Hölger Rauhut. Compressive Sensing and Structured Random Matrices. *Theoretical Foundations and Numerical Methods for Sparse Recovery*, 9, 2010. 163, 165, 169, 171, 172

Hölger Rauhut and Rachel Ward. Sparse legendre expansions via l_1 minimization. *Arxiv preprint arXiv:1003.0251*, 2010. 165

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952. 3, 6, 14, 38, 39, 80, 243

David S. Rosenberg. *Semi-Supervised Learning with Multiple Views*. PhD thesis, University of California, Berkeley, Fall 2008. 181, 193

David S. Rosenberg and Peter L. Bartlett. The rademacher complexity of co-regularized kernel classes. *Proceedings of the Eleventh ICAIS*, 2007. 180, 181, 184, 185, 186, 188, 190, 192

Mark Rudelson and Roman Vershynin. The littlewood-offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008a. 115

Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008b. 167, 168, 173

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *ArXiv e-prints*, mar 2010. 115, 233

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35:395–411, May 2010. 22

Daniil Ryabko and Marcus Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, 405:274–284, October 2008. ISSN 0304-3975. 17, 36, 81, 82, 243

Saburou Saitoh. *Theory of reproducing Kernels and its applications*. Longman Scientific & Technical, Harlow, UK, 1988. 135

Ivan N. Sanov. On the probability of large deviations of random magnitudes. *Matematicheskii Sbornik (N.S.)*, 42(84):11–44, 1957. 109

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE symposium on Foundations of Computer Science.*, FOCS '06, pages 143–152, 2006. 128

Bruno Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. In Fürnkranz and Joachims (2010). xxi, 208

Jürgen Schmidhuber. *Anticipatory Behavior in Adaptive Learning Systems*, chapter Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes, pages 48–76. Springer-Verlag, Berlin, Heidelberg, 2009. ISBN 978-3-642-02564-8. xxi

Bernhard Schölkopf, Ralf Herbrich, and Alexander J. Smola. A generalized representer theorem. pages 416–426, 2001. 183, 185

Paul J. Schweitzer and Abraham Seidmann. Generalized polynomial approximations in markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110: 568–582, 1985. 208

Yevgeny Seldin, Nicolò Cesa-Bianchi, François Laviolette, Peter Auer, John Shawe-Taylor, and Jan Peters. Pac-bayesian analysis of martingales and multiarmed bandits. *ArXiv e-prints*, may 2011a. 121

Yevgeny Seldin, Nicolò Cesa-Bianchi, François Laviolette, Peter Auer, John Shawe-Taylor, and Jan Peters. Pac-bayesian analysis of the exploration-exploitation trade-off. *ArXiv e-prints*, may 2011b. 121

Rocco A. Servedio and Tong Zhang, editors. volume 80 of *COLT '08*, Helsinki, Finland, jul 2008. Omnipress. 257, 260, 264, 269, 275

Shai Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, July 2007. 22, 65

Jennie Si, Andrew G. Barto, Warren B. Powell, and Don Wunsch. *Handbook of Learning and Approximate Dynamic Programming (IEEE Press Series on Computational Intelligence)*. Wiley-IEEE Press, 2004. 208

Olivier Sigaud and Olivier Buffet. *Processus décisionnels de Markov en intelligence artificielle*, volume 1 - principes généraux et applications of *IC2 - informatique et systèmes d'information*. Lavoisier - Hermes Science Publications, 2008. ISBN 978-2746220577. 205

Vikas Sindhwani and David S. Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In Cohen et al. (2008), pages 976–983. ISBN 978-1-60558-205-4. 181, 185, 188, 199

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning*, volume 119 of *ICML '05, ACM International Conference Proceeding Series*, Bonn, Germany, aug 2005. ACM. ISBN 1-59593-180-5. Workshop on Learning with Multiple Views. 181, 184, 185, 193

Alexander J. Smola and Risi I. Kondor. Kernels and regularization on graphs. In *Conference On Learning Theory and 7th Kernel Workshop*, pages 144–158, 2003. 182

Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In Servedio and Zhang (2008), pages 403–414. 180

Ingo Steinwart, Don Hush, and Clint Scovel. A new concentration result for regularized risk minimizers. *IMS Lecture notes monograph series*, 51:260, 2006. 183

Gilles Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences*. Phd thesis, Université Paris-Sud, Orsay, France, May 2005. 24, 29, 68, 81

Gilles Stoltz. *Contributions to the sequential prediction of arbitrary sequences: applications to the theory of repeated games and empirical studies of the performance of the aggregation of experts*. Habilitation à diriger des recherches, Université Paris-Sud, 2011. 29

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michaël L. Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine learning*, ICML '06, pages 881–888, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. 242

Richard S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems 8*, pages 1038–1044. MIT Press, 1996. xxxii, 105, 130

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIP Press, 1998. 205, 208, 209, 210, 227

Richard S. Sutton and Steven D. Whitehead. Online learning with random representations. In *In Proceedings of the 10th International Conference on Machine Learning*, ICML '93, pages 314–321. Morgan Kaufmann, 1993. xxxii, 106, 130

Csaba Szepesvári. Algorithms for reinforcement learning. In Ronald J. Brachman and Thomas G. Dietterich, editors, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, volume 4, pages 1–103. Morgan & Claypool, July 2010. 205

Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer monographs in mathematics. Springer, 2005. 115

Terence Tao, Vai Van, and Manjunath Krishnapur. Random matrices: Universality of esds and the circular law. *Annals of Probability*, 38:2023–2065, 2010. 115

Gerald Tesauro. Temporal difference learning and td-gammon. *Commun. ACM*, 38:58–68, March 1995. ISSN 0001-0782. xvi

Ambuj Tewari and Peter L. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In Platt et al. (2007). 242

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933. xvii, 6

William R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57:450–456, 1935. 6

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994. 125, 164

Andrei N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl 4*, pages 1035–1038, 1963. 125

Alexandre B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the 16th annual Conference On Learning Theory*, pages 303–313, 2003. 141

Sara A. van de Geer. The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007. 164

Sara A. van de Geer and Peter Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. doi: 10.1214/09-EJS506. 164

Vladimir N. Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, jul 2009. 18

Santosh S. Vempala. *The Random Projection Method*. American Mathematical Society, 2004. 226

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In Koller et al. (2008), pages 1729–1736. 21

Christopher J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, 1989. 218

Jason Weston, Christina Leslie, Eugene Ie, Denyong Zhou, Andre Elisseeff, and William S. Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15): 3241–3247, aug 2005. ISSN 1367-4803. 180

Peter Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):143–149, 1980. ISSN 00359246. 25

Eugene P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics (2)*, 67:325–327, 1958. 115

Ronald J. Williams and Leemon C. Baird. Tight performance bounds on greedy policies based on imperfect value functions. In *Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems*, 1994. 208

Christoph Zenger. Sparse grids. In W. Hackbusch, editor, *Parallel Algorithms for Partial Differential Equations, Proceedings of the Sixth GAMM-Seminar*, volume 31 of Notes on Num. Fluid Mechanics, Kiel, 1990. Vieweg-Verlag. 150

Bin Zhao and Changshui Zhang. Compressed spectral clustering. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 344–349, Washington, DC, USA, 2009. IEEE Computer Society. 126

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2563, 2006. 164

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Proceedings of the 17th conference on advances in Neural Information Processing Systems*, NIPS '03, pages 321–328, Vancouver, British Columbia, Canada, dec 2003. MIT Press. ISBN 0-262-20152-6. 180

Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12:349–384, February 2011. 18

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, ICML '03, pages 928–936, 2003. xxx, 22, 63, 64

Index