New insights on concentrations inequalities for martingales applications to statistics and machine learning

Taieb Touati

École Doctorale de Sciences
Mathématiques de Paris Centre

Thèse de Doctorat

en vue de l'obtention du grade de
Docteur ès Sciences de Sorbonne université
Discipline : Mathématiques
Spécialité : Probabilités
présentée par

Taieb Touati

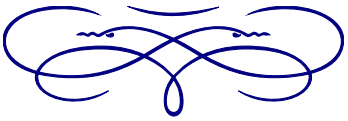
New insights on concentrations inequalities for martingales, applications to statistics and machine learning.

dirigée par Michel Broniatowski
Au vu des rapports établis
par Victor de La Pena et Marianne Clausel
Soutenue publiquement le **22/02/2021 à Paris** devant
le jury composé de :

Vianney Perchet	CREST ENSAE	Examinateur
Alexandra Carpentier	Universität Magdeburg	Examinateur
Gérard Biau	Sorbonne Université	Président du Jury
Michel Broniatowski	Sorbonne Université	Directeur de thèse
Victor De La Pena	Columbia University	Rapporteur
Marianne Clausel	Université de Lorraine	Rapporteur
Bernard Bercu	Université de Bordeaux	Membre invité

Sorbonne Université
Laboratoire de Probabilités, Statistique et Modélisation (LPSM)
4, place Jussieu
75252 Paris Cedex 05

Dédicace



À la mémoire de feu mon père, Prof. ABDERRAHMEN TOUATI ; qui en plus d'avoir été un père exemplaire, demeure une source d'inspiration scientifique.

Remerciements

Mes premiers remerciements s'adressent à mon directeur de thèse Michel Broniatowski à qui j'exprime ma gratitude pour m'avoir accordé une grande liberté dans le choix de mes axes de recherche ce qui m'a permis d'acquérir beaucoup d'autonomie. La thèse fut une expérience très formatrice sur le plan scientifique et une leçon de vie !

Je suis très honoré que Marianne Clausel et Victor De La Peña aient accepté d'être rapporteurs de ma thèse. Je leur suis très reconnaissant du temps qu'ils ont consacré à l'évaluation de ce travail. Je remercie également Alexandra Carpentier, Vianney Perchet et Gérard Biau d'avoir accepté de faire partie de mon jury.
J'exprime d'ailleurs ma profonde gratitude à Gérard, qui m'a conseillé et soutenu tout au long de mon cursus doctoral.

Je tiens à remercier tous mes anciens collègues au LPSM et mes anciens Professeurs à l'ISUP et à l'UPMC. Je remercie Lucien Birgé qui a su nous enseigner avec élégance et rigueur les statistiques mathématiques, un grand merci également à Annick Valibouze pour son dynamisme et son implication au sein de la filère GRIE à l'ISUP, enfin je tiens à exprimer ma reconnaissance pour Philippe Saint-Pierre et Maxime Wolff pour leur écoute et leurs conseils.
Une pensée particulière à Nathalie Akakpo qui en plus d'avoir été une enseignante très pédagogue et bienveillante, m'a initié très jeune à un projet de recherche en collaboration avec Bertrand Iooss de EDF R&D, chercheur senior qui a été un vrai mentor. Je garderai toujours un excellent souvenir des discussions stimulantes avec Qi ming, Félix, Lucie, Dimby, Nicolas, Eric et Nazih, d'ailleurs une collaboration (avec Nazih et Nicolas) en dehors de nos travaux de thèse a été couronnée par

le premier prix d'un Data Challenge organisé par la SFDS. Je remercie également les chercheurs (jeunes et confirmés) avec lesquels j'ai pu échanger durant les divers séminaires et conférences auxquels j'ai pu assisté, deux échanges m'ont particulièrement marqué, ce fut avec Matteo Tacchi et Arnaud LE NY.

Cette thèse n'aurait jamais donné lieu à des résultats pertinents sans l'aide précieuse de Bernard Bercu, qui a fait preuve d'une grande patience et d'un grand courage. Il a honoré une amitié inconditionnelle et sincère. Financièrement, j'ai pu finalisé mes travaux de thèse grâce à un contrat ATER à Paris 1 au SAMM. Je remercie Jean Marc Bardet pour sa confiance et son implication. Je remercie également mes collègues à Paris 1, surtout Caroline Vernier qui m'a beaucoup soutenu tout au long des moments difficiles. Je remercie Paul Doukhan qui m'a permis de présenter mes travaux au sein du prestigieux Institut Henri Poincaré. Cette présentation m'a permis de construire une collaboration originale avec Odalric-Ambrym Maillard, excellent chercheur que je tiens à remercier pour sa disponibilité et sa confiance.

Je remercie la direction de la Maison de Tunisie, en particulier le directeur Tahar Battikh pour son altruisme et son engagement sincère auprès des étudiants. Il a pris en compte ma situation très particulière et j'ai pu grâce à cela continuer mes travaux sereinement durant l'année universitaire 2019-2020.
Une pensée sincère à mes amis : Nafa, Wael, Elyes, Nazih, Amine,Oussama, Youssef Zied, Sodki, Amal, Yasmine, Veronica, Emna.

Je dédie cette thèse à l'ensemble de la famille mathématique en Tunisie et je remercie plus particulièrement : Faouzi Chaabane, Mohamed Ali Jendoubi, Sana Louhichi, Karim Boulabiar, Sadok Kallel et Mohamed Amine Ben Amor.
Enfin mes remerciements vont à toute ma famille en Tunisie et en Algérie, une pensée particulière à ma mère Houria pour son soutien indéfectible, à mon frère et ma belle soeur : Mohammed Yassine et Oumaima pour leur bonne humeur et à ma petite nièce Lina qui a égayé notre quotidien depuis bientôt 5 mois.

J'adresse en dernier lieu un hommage à feu mon père Abderrahmane ; j'espère que de là oú il est, cette thèse le rendra fier et heureux.

Avant-propos

Cette thèse a été financée par un contrat doctoral à Sorbonne Université (ex Université Pierre et Marie Curie), du 1er octobre 2016 au 30 septembre 2019, et réalisée au Laboratoire de Statistique Théorique et Appliquée (LSTA) ensuite Laboratoire de Probabilités, Statistique et Modélisation (LPSM) après fusion avec le Laboratoire de probabilités et modèles aléatoires (LPMA).
La quatrième année de thèse (2019-2020) a été financée par un contrat demi-ATER à Paris 1 Panthéon Sorbonne au laboratoire SAMM. Ladite thèse est composée d'une introduction générale, de trois chapitres et d'une conclusion avec perspectives.

Plan de la thèse

Le Chapitre 1 a été dédié à présenter les outils fondamentaux en lien direct avec les contributions scientifiques de l'auteur. Nous proposons aussi dans ce chapitre une synthèse de l'état de l'art qui sera utile au lecteur non spécialiste. Nous précisons, à toute fin utile que ce n'est pas un chapitre contenant des nouveaux résultats mathématiques mais un résumé basé principalement sur Bercu et al. (2015), Hoi et al. (2018) et le support d'un cours sur les martingales de Prof.Bernard Bercu dispensé au CIRM en 2008.
Nous introduisons en premier lieu, les inégalités de concentrations classiques pour les sommes de variables aléatoires indépendantes, leurs améliorations récentes et les applications qui en découlent. Nous présenterons ensuite, les résultats fondamentaux concernant les martingales et les applications intrinsèquement liées aux statistiques et à l'apprentissage automatique.

Le Chapitre 2 constitue la principale contribution scientifique de l'auteur. Il fait l'objet d'une publication dans **Electronic Communications in Probability**, écrit en collaboration avec Prof. Bernard Bercu (Institut Mathématiques de Bordeaux). Nous introduisons une somme pondérée de la variation quadratique et de la variation quadratique prédictible permettant ainsi d'améliorer, généraliser et d'unifier selon une seule paramétrisation toute une famille d'inégalités de concentrations pour les martingales auto-normalisées. Nous proposons ensuite

des applications en statistiques, notamment pour les processus auto-régressifs non symétriques et les processus de diffusion interne. La grande nouveauté des résultats fournis est leur flexibilité et la possibilité d'obtenir une inégalité optimale pour chaque application. La dernière application est une amélioration d'un des résultats de Cesa-Bianchi ouvrant le champ à des perfectionnement plus substantiels.

Le Chapitre 3 est intrinsèquement lié au chapitre précédent. Ce chapitre constitue une version détaillé d'un preprint élaboré par nos soins en cours de soumission à **Journal of Machine Learning Research**. Nous utilisons les résultats de Bercu and Touati (2019) et des améliorations très fines de l'inégalité de Bernstein Bercu et al. (2015) pour fournir des bornes de risque précises pour un algorithme d'apprentissage séquentiel. Moyennant des simulations numériques, nous montrons l'efficacité de nos bornes et l'amélioration drastique obtenue par rapport aux résultats précédents de Cesa-Bianchi et Gentile. La communauté scientifique de l'apprentissage automatique et séquentiel utilisent exclusivement des inégalités anciennes notamment celles d'Azuma et de Freedman, nos résultats théoriques étant plus performants permettront une amélioration considérable de plusieurs résultats fondamentaux en machine learning (**Non stochastic bandits**,**Region Based Active Learning** et **Aggregation of Experts** entre autres.)

Ces perspectives constituent les objectifs principaux de travaux scientifiques en cours que nous continuerons après la soutenance en collaboration avec Prof. **Odalric-Ambrym Maillard** (INRIA LILLE, Sequel Team) avec lequel nous avons pu échanger sur le sujet suite à une des mes présentations à l'Institut Henri Poincaré. Prof.Maillard nous a posé des questions très pertinentes suite à notre exposé et nous avons pu petit à petit explorer des pistes de perfectionnements des résultats en apprentissage automatique séquentiel en se basant sur les inégalités élaborées dans cette thèse.

Table des matières

Introduction générale

Cette thèse s'intéresse principalement à des inégalités exponentielles pour les martingales réelles à temps discret. Cet objet mathématique a été introduit progressivement entre 1920 et 1940, un article de Crépel (1984) fournit un résumé détaillé et très technique des développements concernant la genèse de ladite théorie.
La mise en oeuvre de cette branche des mathématiques appliquées, coïncide avec l'axiomatisation des probabilités (le fameux Grundbegriffe de Kolmogorov en 1933). Le mathématicien américain Joseph Leo Doob, très largement reconnu dans le milieu mathématique comme le père fondateur des martingales a généralisé les travaux de Paul Lévy, plus précisément Lévy (1937) concernant les sommes de variables aléatoires indépendantes. Jean André Ville fut le premier a fournir une définition mathématique des martingales en ayant pour ambition d'améliorer les travaux de Von Mises (1932). La thèse de Ville (1939) constitue donc la première pierre dans l'édifice de la théorie des martingales.
Nous conseillons au lecteur curieux une excellente biographie de ce pionnier des martingales, faite par Shafer (2009). Les travaux de Ville n'ont pas été très bien accueillis par Paul Lévy qui les estimait dépourvus de rigueur et d'originalité, Mazliak (2009) retrace cela de manière très exhaustive. Il a fallu attendre un congrès de Probabilités à Lyon en 1948 pour que la théorie des martingales soit enfin inscrite comme sujet de recherche pertinent et donnant lieu à des applications intéressantes. L'exposé de Doob est intitulé : **"Application of The Theory of Martingales"**.
Locker (2009) nous fournit des détails de ce congrès, entre autres quelques anecdotes sur Doob, Lévy et Ville. Une de ses remarques fondamentales, c'est que Ville avait abandonné le sujet des martingales et que Lévy n'était plus enclin à y accorder point d'importance. Loin de nous, l'idée de proposer un essai détaillé sur l'épistémologie et l'histoire de la théorie des martingales, cependant nous essayons d'avertir le lecteur qu'une définition qu'il prétendra facile est souvent le fruit d'une dizaine d'années de développements et parfois une suite d'échanges houleux entre mathématiciens.
Sur le plan étymologique, le mot martingale serait issu du provençal martegal, « habitant de Martigues », terme à connotation péjorative au sens d'« extravagant », ensuite désignait une technique de jeux pour éviter de perdre aux jeux de hasard.

Ce terme apprait pour la première fois en 1611 dans dans **"A Dictionarie of the French and English Tongues"** de Randle Cotgrave. (Doob était donc avant-gardiste face aux "Martingales" et probablement doté d'un sens de l'humour assez particulier pour prendre au sérieux un terme relatif aux jeux de hasard, projeté dans le monde mathématique par un jeune docteur !). Cette petite réflexion étant faite, nous passons aux définitions mathématiques formelles.

Tout d'abord, nous allons définir la notion de martingales, présenter quelques résultats de convergence classiques. une description du manuscrit et enfin un résumé des principaux résultats de cette thèse.

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité muni d'une filtration $\mathbb{F} = (\mathcal{F}_n)$ où $\mathbb{F}$ est une suite croissante de sous-tribus de $\mathcal{A}$ et $\mathcal{F}_n$ est la tribu des évènements ayant lieu antérieurement à l'instant n. Une suite (M_n) de variables aléatoires définies sur $(\Omega, \mathcal{A}, \mathbb{P})$, est adaptée à $\mathbb{F}$ si, pour tout $n \geqslant 0$, M_n est $\mathcal{F}_n$-mesurable.

Définition 0.1. Soit (M_n) une suite de variables aléatoires réelles, intégrable et adaptée à $\mathbb{F}$. On dit que (M_n) est une martingale, sous-martingale ou surmartingale à temps discret si, pour tout $n \geqslant 0$, on a respectivement

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n, \qquad \mathbb{E}[M_{n+1}|\mathcal{F}_n] \geqslant M_n, \qquad \mathbb{E}[M_{n+1}|\mathcal{F}_n] \leqslant M_n.$$

Remark 0.2. *Une martingale est constante en terme d'espérance.*

Quelques exemples de martingales :

— Soit (X_n) une suite de variables aléatoires indépendantes et intégrables avec, pour tout $n \geqslant 0$, $\mathbb{E}[X_n] = m$ et soit

$$S_n = \sum_{k=1}^{n} X_k.$$

(S_n) est une martingale, sous-martingale ou surmartingale suivant que $m = 0$, $m \geqslant 0$ ou $m \leqslant 0$, respectivement. La somme de variables aléatoires indépendantes de loi de Rademacher $\mathcal{R}(p)$ avec $0 < p < 1$ est une martingale.

— $P_n = \prod_{k=1}^{n} X_k$. (P_n) est une martingale, sous-martingale ou surmartingale suivant que $m = 1$, $m \geqslant 1$ ou $m \leqslant 1$, respectivement. Le produit de variables aléatoires indépendantes de loi exponentielle $\mathcal{E}(\lambda)$ est une martingale. avec $\lambda > 0$.

Sous certaines conditions, nous pouvons énoncer les équivalents de la loi forte des grands nombres et du théorème de la limite centrale.

Theorem 0.3 (Théorème de Doob)**.** Soit (M_n) une martingale, sous-martingale ou surmartingale, bornée dans $\mathbb{L}^1$, à savoir

$$\sup_{n \geqslant 0} \mathbb{E}[|M_n|] < +\infty.$$

Alors, (M_n) converge presque sûrement vers une variable aléatoire intégrable M_∞.

Theorem 0.4. Soit (M_n) une martingale bornée dans $\mathbb{L}^p$ avec $p \geqslant 1$, à savoir

$$\sup_{n\geqslant 0} \mathbb{E}[|M_n|^p] < +\infty.$$

Si $p > 1$, alors (M_n) converge presque sûrement et dans $\mathbb{L}^p$ vers une variable aléatoire M_∞. Par contre, si $p = 1$, alors (M_n) converge presque sûrement et cette convergence n'a lieu dans $\mathbb{L}^1$ que si (M_n) est équi-intégrable donc, si

$$\lim_{a\to\infty} \sup_{n\geqslant 0} \mathbb{E}[|M_n|\mathrm{I}_{(|M_n|\geqslant a)}] = 0.$$

Une martingale (M_n) est de carré intégrable si, pour tout $n \geqslant 0$, $\mathbb{E}[M_n^2] < +\infty$. Elle est donc bornée dans $\mathbb{L}^2$. Dans ce cas, (M_n^2) est une sous-martingale positive et intégrable.

Définition 0.5. Soit (M_n) une martingale de carré intégrable. On appelle processus croissant associé à (M_n), la suite $(<M>_n)$ définie par $<M>_0= 0$ et, pour tout $n \geqslant 1$,

$$<M>_n= \sum_{k=1}^{n} \mathbb{E}[\Delta M_k^2|\mathcal{F}_{k-1}]$$

avec $\Delta M_n = M_n - M_{n-1}$

Theorem 0.6 (Loi des grands nombres)**.** Soit (M_n) une martingale de carré intégrable et soit $(<M>_n)$ son processus croissant. On pose

$$<M>_\infty= \lim_{n\to\infty} <M>_n .$$

1) Sur $\{<M>_\infty< \infty\}$, (M_n) converge p.s. vers M_∞ de carré intégrable.
2) Sur $\{<M>_\infty= \infty\}$, on a

$$\lim_{n\to\infty} \frac{M_n}{<M>_n} = 0 \qquad \text{p.s.}$$

Remark 0.7. *Une conséquence utile est que si* $<M>_n= \mathcal{O}(a_n)$ *où* (a_n) *est une suite déterministe positive croissante vers l'infini, alors* $M_n = o(a_n)$ *p.s.*

Theorem 0.8 (Théorème limite central)**.** Soit (M_n) une martingale de carré intégrable et soit $(<M>_n)$ son processus croissant. Soit (a_n) une suite déterministe, positive, croissante vers l'infini. On suppose que

1) Il existe une limite déterministe $\ell \geqslant 0$ telle que

$$\frac{<M>_n}{a_n} \xrightarrow{\mathbb{P}} \ell.$$

2) La condition de Lindeberg est satisfaite c'est-à-dire pour tout $\varepsilon > 0$

$$\frac{1}{a_n} \sum_{k=1}^{n} \mathbb{E}[|\Delta M_k|^2 \mathrm{I}_{(|\Delta M_k|\geqslant \varepsilon\sqrt{a_n})}|\mathcal{F}_{k-1}] \xrightarrow{\mathbb{P}} 0.$$

($\xrightarrow{\mathbb{P}}$ désigne la convergence en probabilités.)
Alors, on a

$$\frac{1}{\sqrt{a_n}} M_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \ell). \tag{1}$$

De plus, si $\ell > 0$, on a

$$\sqrt{a_n} \frac{M_n}{<M>_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \ell^{-1}). \tag{2}$$

Nous recommandons au lecteur, des preuves de nature didactique et pédagogique dans les références suivantes : Bercu and Chafaï (2007), Duflo (1997), Hall and Heyde (1980), Mazliak et al. (1998), Neveu (1972), Dacunha-Castelle and Duflo (1985), et Liptser and Shiryayev (1989).

Le but de cette thèse est dans un premier temps, de proposer de nouvelles inégalités concentrations pour les martingales avec des hypothèses minimales très peu contraignantes. Le domaines des inégalités de concentrations est bien établi en probabilités et jouit d'un essor particulier depuis deux décennies en particulier pour la théorie des martingales. Nous élaborerons dans le premier chapitre un état de l'art exhaustif des inégalités de concentrations pour les sommes de variables aléatoires indépendantes et pour les martingales. Ensuite nous présenterons succinctement le lien entre cette théorie et le domaine de l'apprentissage automatique séquentiel.

Le chapitre 2 est dédié à la mise en oeuvre de nouvelles inégalités de concentration pour les martingales auto-normalisées. Ces résultats généralisent, perfectionnent ceux de la littérature tout en permettant l'unification des inégalités à l'aide d'une paramétrisation judicieuse. Nous mettons en évidence des applications en statistiques et en apprentissage automatique.

Enfin à l'aide des résultats théorique établis tout au long du chapitre 2 et de certaines inégalités mentionnées dans l'état de l'art, nous améliorons les résultats de Cesa-Bianchi and Gentile (2008). Ce perfectionnement des bornes de risques est mis en valeur à l'aide de simulations numériques diverses. Ce protocole d'amélioration en utilisant des outils probabilistes plus performants peut s'appliquer à d'autres cas de figure en apprentissage automatique ouvrant ainsi tout un champ de recherche à explorer.

Chapitre 1

State of the art

Abstract

This chapter introduces the mathematical tools necessary for the entire manuscript and provides a review of the literature. We present at the beginning, the classical inequalities for the sums of independent random variables, then we present the key results concerning martingales and autonormalized martingales. Finally we provide some applications in statistics and machine learning.

1.1 Concentration inequalities for sums

We present in this section, the classical inequalities for the sums of independent random variables. We also provide some more efficient improvements that will allow us to improve the accuracy of risk tail bounds for online learning algorithms among others. This type of inequality is closely linked to applications in inferential statistics.

1.1.1 Bernstein's inequality for martingales

Bernstein (1927) was a pioneer of exponential inequalities for sums of independent random variables. We owe him the following theorem, which is most often used as the gold standard for controlling a deviation.
This is far from optimal, since we can considerably improve this inequality and obtain smaller exponential bounds.

Theorem 1.1. Let $X_1, \ldots, X_n$ be a finite sequence of independent random variables with finite variances. Denote

$$S_n = X_1 + \cdots + X_n, \qquad \mathcal{V}_n = \mathbb{E}[X_1^2] + \cdots + \mathbb{E}[X_n^2], \qquad v_n = \frac{\mathcal{V}_n}{n}. \tag{1.1}$$

Assume that $\mathbb{E}[S_n] = 0$ and that there exists some positive constant c such that, for any integer $p \geqslant 3$,

$$\sum_{k=1}^{n} \mathbb{E}\Big[|X_k|\Big]^p \leqslant \frac{p! c^{p-2}}{2} \mathcal{V}_n. \tag{1.2}$$

Then, for any positive x,

$$\mathbb{P}(S_n \geqslant nx) \leqslant \exp\Big(-\tfrac{nx^2}{2(v_n + cx)}\Big). \tag{1.3}$$

In addition, we also have, for any positive x,

$$\mathbb{P}\Big(S_n > n(cx + \sqrt{2v_n x})\Big) \leqslant \exp(-nx). \tag{1.4}$$

Remark 1.2. *It is not necessary to assume that the random variables $X_1, \ldots, X_n$ are centered. We only have to suppose that $\mathbb{E}[S_n] = 0$. In the centered case, $\mathcal{V}_n$ coincides with $V_n = Var(S_n)$. Otherwise, $\mathcal{V}_n$ is obviously larger that V_n.*

1.1.2 Hoeffding's inequality

In this subsection, the focus is set on the classical Hoeffding's inequality Hoeffding (1963), where it is necessary for the random variables to be bounded from above and below. We shall also establish Antonov's type extensions to this inequality. Starting with Hoeffding's inequalitiy, The proof relies on the following lemmas which provide upper bounds for the variances and the Laplace transformations of the random variables $X_1, \ldots, X_n$.

Theorem 1.3 (Hoeffding's inequality)**.** Let $X_1, \ldots, X_n$ be a finite sequence of independent random variables. Assume that for all $1 \leqslant k \leqslant n$, one can find two constants $a_k < b_k$ such that $a_k \leqslant X_k \leqslant b_k$ almost surely. Denote $S_n = X_1 + \ldots + X_n$. Then, for any positive x,

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geqslant x) \leqslant 2\exp\Big(-\frac{2x^2}{D_n}\Big) \quad \text{where} \quad D_n = \sum_{k=1}^{n} (b_k - a_k)^2. \tag{1.5}$$

The proof of Hoeffding's inequalitiy relies on the following lemmas which give upper bounds for the variances and the Laplace transforms of the random variables $X_1, \ldots, X_n$.

Lemma 1.4. *Let X be a random variable with finite variance σ^2. Assume that $a \leqslant X \leqslant b$ almost surely for some real constants a and b. Denote $m = \mathbb{E}[X]$. Then,*

$$\sigma^2 \leqslant (b - m)(m - a) \leqslant \frac{(b - a)^2}{4}. \tag{1.6}$$

Démonstration. The convexity of the square function implies that $X^2 \leqslant (a + b)X - ab$ almost surely. Hence $\sigma^2 = \mathbb{E}[X^2] - m^2 \leqslant (a + b)m - ab - m^2 \leqslant -ab + (a + b)^2/4$, which implies the lemma. □

□

Lemma 1.5. *Let X be a random variable with finite variance σ^2. Assume that $a \leqslant X \leqslant b$ almost surely for some real constants a and b. Then, for any real t,*

$$\log\Big(\mathbb{E}[\exp(tX)]\Big) \leqslant t\mathbb{E}[X] + \frac{t^2}{8}\Big(b-a\Big)^2.$$

Démonstration. Let L and ℓ be the Laplace and log-Laplace transforms of X. As the random variable X is bounded from above and from below, L and ℓ are real analytic functions. Moreover, for any real t, $\ell(t) = \log L(t)$,

$$\ell'(t) = \frac{L'(t)}{L(t)} \qquad \text{and} \qquad \ell''(t) = \frac{L''(t)}{L(t)} - \left(\frac{L'(t)}{L(t)}\right)^2.$$

Consider the classical change of probability

$$\frac{d\mathbb{P}_t}{d\mathbb{P}} = \exp\Big(tX - \ell(t)\Big) = \frac{\exp(tX)}{L(t)}$$

and denote by $\mathbb{E}_t$ the expectation associated with $\mathbb{P}_t$. One can observe that for any integrable random variable Y,

$$\mathbb{E}_t[Y] = \frac{\mathbb{E}[Y\exp(tX)]}{L(t)}.$$

In particular,

$$E_t[X] = \frac{\mathbb{E}[X\exp(tX)]}{L(t)} = \frac{L'(t)}{L(t)}, \quad E_t[X^2] = \frac{\mathbb{E}[X^2\exp(tX)]}{L(t)} = \frac{L''(t)}{L(t)}.$$

Consequently, $\ell''(t) = \mathbb{E}_t[X^2] - \mathbb{E}_t^2[X]$, which means that $\ell''(t)$ is equal to the variance of the random variable X under the new probability $\mathbb{P}_t$. As the random variable X takes its values in $[a,b]$ almost surely, we may apply Lemma 1.4 under the new probability $\mathbb{P}_t$, which gives $\ell''(t) \leqslant (b-a)^2/4$. Since $\ell(0) = 0$ and $\ell'(0) = \mathbb{E}[X]$, it completes the proof of Lemma 1.5. □

□

Proof of Theorem 1.3. We shall now proceed to the proof of Hoeffding's inequality. We deduce from Lemma 1.5 together with the independence of the random variables $X_1, \ldots, X_n$ that, for any real t,

$$\log \mathbb{E}\Big[\exp(tS_n)\Big] = \sum_{k=1}^{n} \log \mathbb{E}\Big[\exp(tX_k)\Big] \leqslant t\mathbb{E}[S_n] + \frac{t^2}{8}D_n \tag{1.7}$$

where D_n is given by (1.5). For any positive t, it follows from Markov's inequality applied to $\exp(tS_n)$ that

$$\log \mathbb{P}(S_n \geqslant \mathbb{E}[S_n] + x) \leqslant -tx - t\mathbb{E}[S_n] + \log \mathbb{E}[\exp(tS_n)]. \tag{1.8}$$

Consequently, inequalities (1.8) and (1.7) imply that for all $x \geqslant 0$ and $t > 0$,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant x) \leqslant \exp\left(-tx + \frac{t^2}{8} D_n\right).$$

By taking the optimal value $t = 4x/D_n$, we find that

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant x) \leqslant \exp\left(-\frac{2x^2}{D_n}\right). \tag{1.9}$$

Replacing X_k by $-X_k$, we obtain by the same token that for all $x \geqslant 0$,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \leqslant -x) \leqslant \exp\left(-\frac{2x^2}{D_n}\right). \tag{1.10}$$

Therefore, (1.5) follows from (1.9) and (1.10), which completes the proof of Theorem 1.3. □

1.1.3 Binomial rate functions

In this subsection, the assumption is that $X_1, \ldots, X_n$ is a finite sequence of independent random variables bounded from above. More precisely, the assumption establishes the ground basis that there exists a positive constant b such that, for all $1 \leqslant k \leqslant n$,

$$X_k \leqslant b \qquad \text{a.s.} \tag{1.11}$$

. The following results outperform those of Bernstein, which will prove to be very useful in the rest of the thesis, more precisely in Chapitre 3.

Theorem 1.6. Let $X_1, \ldots, X_n$ be a finite sequence of independent random variables with finite variances satisfying (1.11) for some positive real b. Let S_n and v_n by defined as in (1.1) and assume that $\mathbb{E}[S_n] = 0$. Then, for any $v \geqslant v_n$ and for any x in $[0, b]$,

$$\begin{aligned} \mathbb{P}(S_n \geqslant nx) &\leqslant \exp\Big(-n\Big(\Big(\frac{v+bx}{v+b^2}\Big)\log\Big(1+\frac{bx}{v}\Big) + \Big(\frac{b^2-bx}{b^2+v}\Big)\log\Big(1-\frac{x}{b}\Big)\Big)\Big), \\ &\leqslant \exp\Big(-ng(b,v)x^2\Big), \end{aligned} \tag{1.12}$$

where

$$g(b,v) = \begin{cases} \dfrac{b^2}{(b^4-v^2)}\log\Big(\dfrac{b^2}{v}\Big) & \text{if} \quad v < b^2, \\ \dfrac{1}{2v} & \text{if} \quad v \geqslant b^2. \end{cases} \tag{1.13}$$

If we apply **Theorem 1.6** to independent random variables with values in $[0,1]$. We obtain exactly **Theorem 1** in Hoeffding (1963).

Theorem 1.7. Let $X_1,\dots,X_n$ be a finite sequence of independent random variables with values in $[0,1]$ and denote $\mu=\mathbb{E}[S_n]/n$. Then, for any x in $]\mu,1[$,

$$\begin{aligned}\mathbb{P}(S_n\geqslant nx) &\leqslant \exp\Big(-n\Big(x\log\Big(\frac{x}{\mu}\Big)+(1-x)\log\Big(\frac{1-x}{1-\mu}\Big)\Big)\Big),\\ &\leqslant \exp\Big(-ng(\mu)(x-\mu)^2\Big), \qquad (1.14)\\ &\leqslant \exp\Big(-2n(x-\mu)^2\Big),\end{aligned}$$

where

$$g(\mu)=\begin{cases}\dfrac{1}{1-2\mu}\log\Big(\dfrac{1-\mu}{\mu}\Big) & \text{if} \quad 0<\mu<\dfrac{1}{2},\\[2ex] \dfrac{1}{2\mu(1-\mu)} & \text{if} \quad \dfrac{1}{2}\leqslant\mu<1.\end{cases}$$

1.1.4 Bennett's inequality

In this subsection, we deduce Bennett (1962)'s type inequalities from the results of subsection 1.1.3. First of all, let h and h_w be the functions defined by

$$h(x)=\begin{cases}(1+x)\log(1+x)-x & \text{if} \quad x>-1,\\ 1 & \text{if} \quad x=-1,\\ +\infty & \text{if} \quad x<-1,\end{cases} \qquad (1.15)$$

and

$$h_w(x)=\begin{cases}\dfrac{h(wx)}{w^2} & \text{if} \quad w\neq 0,\\[2ex] \dfrac{x^2}{2} & \text{if} \quad w=0.\end{cases} \qquad (1.16)$$

Theorem 1.8. Let $X_1,\dots,X_n$ be a finite sequence of independent random variables satisfying (1.11) for some positive constant b. Let S_n and v_n by defined as in (1.1) and assume that $\mathbb{E}[S_n]=0$. Let $w_n=(b/v_n)-(1/b)$. Then, for any x in $[0,b]$,

$$\begin{aligned}\mathbb{P}(S_n\geqslant nx) &\leqslant \exp\Big(-\frac{n}{v_n}h_{w_n}(x)\Big), \qquad (1.17)\\ &\leqslant \exp\Big(-\frac{nv_n}{b^2}h\Big(\frac{bx}{v_n}\Big)\Big),\end{aligned}$$

where the above functions are given by (1.15) and (1.16). Hence, if $v_n\geqslant b^2$, then, for any positive x,

$$\mathbb{P}(S_n\geqslant nx)\leqslant\exp\Big(-\frac{nx^2}{2v_n}\Big). \qquad (1.18)$$

Furthermore, for any x in $[0, b]$,

$$\mathbb{P}(S_n \geqslant nx) \leqslant \exp\left(-\frac{nx^2}{2(v_n + bx/3)}\right). \tag{1.19}$$

1.2 Concentration inequalities for martingales

This section is devoted to concentration inequalities for martingales such as Azuma-Hoeffding, Freedman and De la Peña inequalities. This type of inequality serves as a theoretical basis of wich we will improve some foundations in the next chapter. These probabilistic tools constitute a very effective means of improving the non-asymptotic control of deviations in statistics and machine learning.

1.2.1 Azuma-Hoeffding inequalities

Throughout this subsection, (M_n) is a square integrable martingale with bounded differences, adapted to a filtration $\mathbb{F} = (\mathcal{F}_n)$, such that $M_0 = 0$. Its increasing process is defined by $<M>_0 = 0$ and, for all $n \geqslant 1$,

$$<M>_n = \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}]. \tag{1.20}$$

In all the sequel, we shall denote $\Delta M_n = M_n - M_{n-1}$ and

$$V_n = <M>_n - <M>_{n-1} = \mathbb{E}[(M_n - M_{n-1})^2 | \mathcal{F}_{n-1}]. \tag{1.21}$$

Hoeffding (1963) realized that Theorem 1.3 holds also true for martingales (*see* Hoeffding (1963), p. 18). More precisely, assume that (M_n) is a martingale such that, for all $1 \leqslant k \leqslant n$, one can find two constants $a_k < b_k$ satisfying $a_k \leqslant \Delta M_k \leqslant b_k$ almost surely. Then, for any positive x,

$$\mathbb{P}(M_n \geqslant x) \leqslant \exp\left(-\frac{2x^2}{D_n}\right) \quad \text{where} \quad D_n = \sum_{k=1}^{n} (b_k - a_k)^2. \tag{1.22}$$

Later, Azuma (1967) gave a complete proof of (1.22) in the symmetric case $a_k = -b_k$. Our goal in the section is to provide new versions of Azuma-Hoeffding type inequalities. However, instead of considering deterministic bounds on the increments, we will only assume that, for all $1 \leqslant k \leqslant n$, one can find a negative bounded random variable A_k and a positive bounded random variable B_k such that the couple (A_k, B_k) is $\mathcal{F}_{k-1}$-measurable and it satisfies

$$A_k \leqslant \Delta M_k \leqslant B_k \qquad \text{a.s.} \tag{1.23}$$

Our strategy is motivated by the widespread example of martingale transforms given below. Let (ε_n) be a sequence of random variables, adapted of $\mathbb{F}$, such that for all $k \geqslant 1$, $\mathbb{E}[\varepsilon_k | \mathcal{F}_{k-1}] = 0$ and $a_k \leqslant \varepsilon_k \leqslant b_k$ almost surely, for two constants

$a_k < b_k$. In addition, let (ϕ_n) be a sequence of positive and bounded random variables, adapted of $\mathbb{F}$. For all $n \geqslant 1$, denote

$$M_n = \sum_{k=1}^{n} \phi_{k-1}\varepsilon_k. \tag{1.24}$$

The sequence (M_n) is commonly called a martingale transform. One can obviously see that (1.23) holds true with $A_k = a_k\phi_{k-1}$ and $B_k = b_k\phi_{k-1}$. In Subsection 1.2.1, we extend the so-called Kearns and Saul (1998)'s inequality to martingales with differences bounded from above. Subsection 1.2.1 is devoted to martingales satisfying symmetric boundedness assumptions, while Subsection 1.2.1 deals with several improvements of Azuma-Hoeffding's inequality, including van de Geer (2002)'s inequality.

Martingales with differences bounded from above

Throughout this subsection, we assume that (M_n) is a martingale satisfying, for all $1 \leqslant k \leqslant n$, the one-sided boundedness condition

$$\Delta M_k \leqslant B_k \qquad \text{a.s.} \tag{1.25}$$

where B_k is a positive and bounded $\mathcal{F}_{k-1}$-measurable random variable. 79

Under this assumption, qe the following results.

Theorem 1.9. Let (M_n) be a square integrable martingale satisfying (1.25) and let (V_n) be the sequence given by (1.21). Denote by φ the function

$$\varphi(v) = \begin{cases} \dfrac{1-v^2}{|\log(v)|} & \text{if} \quad v < 1, \\ 2v & \text{if} \quad v \geqslant 1. \end{cases} \tag{1.26}$$

Then, for any positive x and y,

$$\mathbb{P}(M_n \geqslant x, \mathcal{A}_n \leqslant y) \leqslant \exp\left(-\frac{x^2}{y}\right) \quad \text{where} \quad \mathcal{A}_n = \sum_{k=1}^{n} B_k^2 \varphi\left(\frac{V_k}{B_k^2}\right). \tag{1.27}$$

Consequently,

$$\mathbb{P}(M_n \geqslant x, 6 <M>_n + \mathcal{C}_n \leqslant y) \leqslant \exp\left(-\frac{3x^2}{y}\right) \tag{1.28}$$

where

$$\mathcal{C}_n = \sum_{k=1}^{n} \left(B_k - \frac{V_k}{B_k}\right)_+^2.$$

Consequently, for any positive x,

$$\mathbb{P}(M_n \geqslant x) \leqslant \exp\left(-\frac{x^2}{\|\mathcal{A}_n\|_\infty}\right) \leqslant \exp\left(-\frac{3x^2}{\|6 <M>_n + \mathcal{C}_n\|_\infty}\right). \tag{1.29}$$

The Following lemma provides a computationally efficient upper bound for the function φ.

Lemma 1.10. *For any v in $]0,1]$,*

$$\varphi(v) \leqslant \frac{1}{3}(1+4v+v^2). \tag{1.30}$$

Symmetric conditions for bounded difference martingales

This subsection deals with the situation where the martingale (M_n) satisfies, for all $1 \leqslant k \leqslant n$, the symmetric boundedness condition

$$|\Delta M_k| \leqslant B_k \qquad \text{a.s.} \tag{1.31}$$

where B_k is a positive and bounded $\mathcal{F}_{k-1}$-measurable random variable. It is inspired by the original work of Azuma (1967).

Theorem 1.11. Let (M_n) be a square integrable martingale such that $M_0 = 0$. Assume that (M_n) satisfies (1.31). Then, for any positive x and y,

$$\mathbb{P}(M_n \geqslant x, 5 <M>_n + \mathcal{B}_n \leqslant y) \leqslant \exp\left(-\frac{3x^2}{y}\right) \tag{1.32}$$

where

$$\mathcal{B}_n = \sum_{k=1}^{n} B_k^2.$$

Consequently, for any positive x,

$$\mathbb{P}(M_n \geqslant x) \leqslant \exp\left(-\frac{3x^2}{\|5 <M>_n + \mathcal{B}_n\|_\infty}\right). \tag{1.33}$$

Asymmetric conditions for bounded difference martingales

We now focus our attention on asymmetric boundedness conditions. As in van de Geer (2002), we assume that (M_n) satisfies, for all $1 \leqslant k \leqslant n$, the asymmetric boundedness condition

$$A_k \leqslant \Delta M_k \leqslant B_k \qquad \text{a.s.} \tag{1.34}$$

where the couple (A_k, B_k) is $\mathcal{F}_{k-1}$-measurable and A_k is a negative and bounded random variable, while B_k is a positive and bounded random variable.

Theorem 1.12. Let (M_n) be a square integrable martingale such that $M_0 = 0$. Assume that (M_n) satisfies (1.34). Then, for any positive x and y,

$$\mathbb{P}(M_n \geqslant x, 2 <M>_n + \mathcal{D}_n \leqslant y) \leqslant \exp\left(-\frac{3x^2}{y}\right) \tag{1.35}$$

where

$$\mathcal{D}_n = \sum_{k=1}^{n} (B_k - A_k)^2.$$

Consequently, for any positive x,

$$\mathbb{P}(M_n \geqslant x) \leqslant \exp\left(-\frac{3x^2}{\|2 <M>_n + \mathcal{D}_n\|_\infty}\right). \tag{1.36}$$

1.2.2 Bernstein's inequality for martingales

Bernstein (1937) extended his exponential inequalities for sums to martingale differences. More precisely, he obtained an extension of inequality (1.3) under the condition that for integer $p \geqslant 3$ and for all $1 \leqslant k \leqslant n$,

$$\mathbb{E}\Big[|\Delta M_k|^p | \mathcal{F}_{k-1}\Big] \leqslant \frac{p! c^{p-2}}{2} V_k \qquad \text{a.s.} \tag{1.37}$$

where (V_n) is the sequence given by (1.21). In this subsection, we will focus our attention on the following improvement of Bernstein (1937)'s inequality.

Theorem 1.13. Let (M_n) be a square-integrable martingale such that $M_0 = 0$. Then, for any positive x and for any positive y,

$$\mathbb{P}(M_n \geqslant nx, <M>_n \leqslant ny) \leqslant \exp\Big(-\tfrac{nx^2}{2(y+cx)}\Big). \tag{1.38}$$

In addition, we also have, for any positive x and for any positive y,

$$\mathbb{P}\Big(M_n > n(cx + \sqrt{2xy}), <M>_n \leqslant ny\Big) \leqslant \exp(-nx). \tag{1.39}$$

1.2.3 De la Pena's inequalities

In order to avoid any boundeness or moment assumption, De la Peña (1999) proposes new exponential inequalities in the particular case where (M_n) is a conditionally symmetric martingale. It involves its total quadratic variation

$$[M]_n = \sum_{k=1}^{n} \Delta M_k^2.$$

1.2.4 Conditionally symmetric martingales

Définition 1.14. Let (M_n) be a martingale adapted to a filtration $\mathbb{F} = (\mathcal{F}_n)$. We shall say that (M_n) is conditionally symmetric if, for all $n \geqslant 1$, the distribution of its increments ΔM_n given $\mathcal{F}_{n-1}$ is symmetric.

Theorem 1.15. Let (M_n) be a square integrable and conditionally symmetric martingale such that $M_0 = 0$. Then, for any positive x and for any positive y,

$$\mathbb{P}(M_n \geqslant x, [M]_n \leqslant y) \leqslant \exp\Big(-\frac{x^2}{2y}\Big). \tag{1.40}$$

For self-normalized martingales, the result is as follows.

Theorem 1.16. Let (M_n) be a square integrable and conditionally symmetric martingale, such that $M_0 = 0$. Then, for any positive x and y, and for all $a \geqslant 0$, $b > 0$

$$\mathbb{P}\Big(\frac{M_n}{a + b[M]_n} \geqslant x\Big) \leqslant \sqrt{\mathbb{E}\Big[\exp\Big(-\frac{x^2}{2}\Big(2ab + b^2[M]_n\Big)\Big)\Big]}, \tag{1.41}$$

$$\mathbb{P}\Big(\frac{M_n}{a + b[M]_n} \geqslant x, [M]_n \geqslant y\Big) \leqslant \exp\Big(-\frac{x^2}{2}\Big(2ab + b^2 y\Big)\Big). \tag{1.42}$$

Thanks to the notion of heavy martingales on the left, respectively on the right and the control of both the total quadratic variation $[M]_n$ and the increasing process $< M >_n$, Bercu and Touati (2008) have further relaxed the assumptions made by De la Peña (1999) while improving the accuracy of exponential bounds. Delyon (2009) generalize these inequalities for the sums of weakly dependent random variables, he provides an improvement of the main result of Bercu and Touati (2008) regarding martingales.
We will standardize and generalize this type of inequality according to a tailor-made parameterization which will improve the precision of these said inequalities while guaranteeing an optimal bound for each application considered.
In the book of Bercu et al. (2015), several applications of the theory of concentration inequalities in the fields of applied probability and statistics are provided in particular : Autoregressive process, Random permutations and Random matrices.
We will focus on the impact of this theory on the very promising field anchored in the present, that of online learning.

1.3 Pedagogical survey on online learning

In this section, we present the field of online learning in contrast to classic machine learning. Based on a key example of binary classification in an online manner, we present a panoply of results that we will improve in chapter 3 of the thesis. We highlight the appearance of a martingale structure to make the link with the inequalities of concentrations.

We advise the reader wishing more details the following reference Hoi et al. (2018) which constitutes a major bibliographic study : very accessible and extremely clear.

1.3.1 Batch Learning vs Online Learning

The traditional machine learning paradigm often works in batches i.e the number of training samples. Such a paradigm is restrictive because it requires that all the training data are available to achieve the learning task.
Batch learning is often considered as the main concept around which numerous cutting-edge machine learning models, including deep networks, are built. In a stochastic environment where access to a finite training set is usually granted, an example is sampled at random from the latter at every step in time. The model compares the prediction generated using the picked example to the ground truth in order to update its parameters with the error gradient. Usually, each element of the training set goes through the same process repeatedly for several epochs until a (heuristic) stopping criterion is met.

Both Batch machine learning and online machine learning share the same mathematical foundations which is the reason why both concepts are rarely distinguished from each other. However, It is important to argue that online machine learning overpowers stochastic gradient descent mainly for the reason that overfitting, convergence, and epochs are not relevant in an online machine learning paradigm.

1.3.2 Online Supervised Learning and related theory

Without loss of generality, We will base all our analysis on a classic online learning problem, i.e., online binary classification. We then present the theories to which this problem is intrinsically connected to know : statistical learning theory, online convex optimization and game theory. These areas constitute the theoretical foundations of the field for online learning. The book of Cesa-Bianchi and Lugosi (2006) highlights these aspects in detail.

Consider an online binary classification task ; On each round, a learner receives a data instance, and then provide a prediction of the instance. After making the prediction, the learner acquires the true output from the environment as a feedback. Based on the feedback, the learner can measure the loss suffered, depending on the difference between the prediction and the true outuput. Finally, the learner updates its prediction model by some strategy so as to enhance the predictive accuracy.
More formally, We call hypothesis H, the classifier or regressor generated by a learning algorithm after training. The predictive performance of hypothesis H is measured by the theoretical risk denoted $R(H)$, which is the expected loss on a

realisation $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ drawn from the underlying distribution

$$R(H) = \mathbb{E}[\ell(H(X), Y)]$$

where ℓ is a nonnegative and bounded loss function. Denote by $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ a training data set of independent random variables sharing the same unknown distribution as (X, Y). Our goal is to predict $Y_{n+1} \in \{-1, 1\}$ given $X_{n+1} \in \mathbb{R}^d$, on the basis of $\mathcal{S}_n$. Let $\mathcal{H}_n = \{H_0, H_1, \ldots, H_{n-1}\}$ be a finite ensemble of hypotheses generated by an online learning algorithm where the initial hypothesis H_0 is arbitrarily chosen.The empirical risk and the average risk associated with the ensemble of hypotheses $\mathcal{H}_n$ and the training data set $\mathcal{S}_n$ are respectively given by

$$\widehat{R}_n = \frac{1}{n}\sum_{k=1}^{n} \ell(H_{k-1}(X_k), Y_k) \qquad \text{and} \qquad R_n = \frac{1}{n}\sum_{k=1}^{n} R(H_{k-1}). \tag{1.43}$$

.
We can interpret the hypotheses as being weights that will be updated at each step of the learning.

Statistical Learning Theory

Assume instance $\mathbf{x}_t$ is generated randomly from an unknown distribution $P(\mathbf{X})$ and its output label y is also generated with an unknown distribution $P(y \mid \mathbf{X})$. The joint distribution of labeled data is $P(\mathbf{X}, y) = P(\mathbf{X})P(y \mid \mathbf{X})$. The aim of a learning problem as defined by Vapnik (1999),is to find a prediction function $f(\mathbf{x})$ that minimizes the expected value of the loss function :

$$R(f) = \int \ell(y, f(\mathbf{x}))\mathrm{d}P(x, y)$$

which is also termed as the True Risk function. The solution $f^* = \arg\min R(f)$ is the optimal predictor. In general, the true risk function $R(f)$ cannot be computed directly because of the unknown distribution $P(x, y)$. In practice, we approximate the true risk by estimating the risk over a finite collection of instances $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ drawn i.i.d., which is called the "Empirical Risk" or "Empirical Error"

$$R_{emp}(f) = \frac{1}{T}\sum_{k=1}^{T} \ell\left(y_k, f\left(\mathbf{x}_k\right)\right)$$

The problem of learning via the Empirical Error Minimization (ERM) is to find a hypothesis f over a hypothesis space $\mathcal{F}$ by minimizing the Empirical Error :

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} R_{emp}(f)$$

ERM is the theoretical base for many machine learning algorithms. For example, in the problem of binary classification, when assuming $\mathcal{F}$ is the set of linear

classifiers and the hinge loss is used, the ERM principle indicates that the best linear model w can be trained by minimizing the following objective

$$R_{emp}(\mathbf{w}) = \frac{1}{T}\sum_{k=1}^{T} \max\left(0, 1 - y_k \mathbf{w}^\top \mathbf{x}_k\right)$$

Thereby, an online learning problem is naturally interpreted as a statistical learning problem in the theory of Vapnik (1999).

Convex Optimization Theory

Several online learning problems can obviously be formulated as an Online Convex Optimization (OCO) task.

An online convex optimization task, mainly consists of two elements : a convex set $\mathcal{S}$ and a convex cost function $\ell_t(\cdot)$. At each time step t, the online algorithm pick a weight vector $\mathbf{w}_t \in \mathcal{S}$; after that, it suffers a loss $\ell_t(\mathbf{w}_t)$, which is computed based on a convex cost function $\ell_t(\cdot)$ defined over $\mathcal{S}$. The aim of the online algorithm is to choose a sequence of hypothesis $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T$ such that the regret in hindsight can be minimized. More formally, an online algorithm aims to reach the smallest possible regret R_T after T rounds, where the regret R_T is defined as :

$$R_T = \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \inf_{\mathbf{w}^* \in \mathcal{S}} \sum_{t=1}^{T} \ell_t(\mathbf{w}^*)$$

where $\mathbf{w}^*$ is the solution that minimizes the convex objective function $\sum_{t=1}^{T} \ell_t(\mathbf{w})$ over $\mathcal{S}$.

As an example, we consider an online binary classification task with online Support Vector Machines (SVM). We can define the loss function $\ell(\cdot)$ as $\ell_t(\mathbf{w}_t) = \max(0, 1 - y_t \mathbf{w}^\top \mathbf{x})$ and the convex set S as $\left\{\forall \mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\| \leqslant C\right\}$ for some constant parameter C. Various algorithms can solve this problem. We suggest to the reader, some easy to access and fairly educational sources Shalev-Shwartz et al. (2011) and Hazan (2016).

Game Theory

We will show in this subsubsection that Game theory is intrinsically related to online learning. An online prediction task can be formulated as a problem of learning to play a repeated game between a learner and an environment Freund and Schapire (1999)

As an illustrative example we will consider binary online classification. At each step, the algorithm pick one class from a finite number of classes and the environment reveals the true class label. We assume that the environment is stable. The algorithm aims to carry out as well as the best fixed strategy. The classic

online classification problem thus can be modeled by the game theory under the simplest assumption, full feedback and a stable environment.

1.3.3 Online learning and concentrations inequalities

Based on (1.43), the following quantity

$$M_n = \sum_{k=1}^{n} \Big(R(H_{k-1}) - \ell(H_{k-1}(X_k), Y_k) \Big), \tag{1.44}$$

is interpreted both as an square integrable martingale and as a sum of bounded independent random variables. We can therefore use the inequalities mentioned in the previous sections to control the deviation of M_n i.e $\mathbb{P}\Big(\frac{M_n}{n} \geqslant x \Big)(x \in [0,1])$. $\frac{M_n}{n}$ is interpreted statistically as the difference between the theoretical risk and the empirical risk for an online algorithm.
effectively controlling this amount is the basis for setting up efficient and very precise risk tail bounds.

Aside from the work of Rakhlin and Sridharan (2017), the overwhelming majority of the scientific community in the field of online learning uses classic concentration inequalities for the implementation of risk tail bounds.
We set ourselves as objective in this thesis, to provide a substantial improvement for risk tail bounds of an arbitrary online learning algorithm based on new concentrations inequalities that we establish in the next chapter.

Chapitre 2

New insights on concentration inequalities for self-normalized martingales

This chapter is an extended version of the article Bercu and Touati (2019) published in Electronic communications in Probabilty.

Abstract

We propose new concentration inequalities for self-normalized martingales. The main idea is to introduce a suitable weighted sum of the predictable quadratic variation and the total quadratic variation of the martingale. It offers much more flexibility and allows us to improve previous concentration inequalities. Statistical applications on autoregressive process, internal diffusion-limited aggregation process, and online statistical learning are also provided. We develop some numerical simulations to compare the bounds that we have developed compared to that of the literature.

2.1 Introduction

Let (M_n) be a locally square integrable real martingale adapted to a filtration $\mathbb{F} = (\mathcal{F}_n)$ with $M_0 = 0$. The predictable quadratic variation and the total quadratic variation of (M_n) are respectively given by

$$<M>_n = \sum_{k=1}^{n} \mathbb{E}[\Delta M_k^2 | \mathcal{F}_{k-1}] \qquad \text{and} \qquad [M]_n = \sum_{k=1}^{n} \Delta M_k^2$$

where $\Delta M_n = M_n - M_{n-1}$ with $<M>_0 = 0$ and $[M]_0 = 0$. Since the pioneer work ofAzuma (1967), Hoeffding (1963), a wide literature is available on concentration inequalities for martingales. We refer the reader to the recent books

Bercu et al. (2015), Boucheron et al. (2013), De la Peña et al. (2007) where the celebrated Azuma-Hoeffding, Freedman, Bernstein, and De la Peña inequalities are provided. Over the last two decades, there has been a renewed interest in this area of probability. More precisely, extensive studies have been made in order to establish concentration inequalities for (M_n) without boundedness assumptions on its increments Bercu and Touati (2008), Delyon (2009), Fan et al. (2015), Pinelis (2014), Rio (2013). For example, it was established in Bercu and Touati (2008) that for any positive x and y,

$$\mathbb{P}(|M_n| \geqslant x, [M]_n + <M>_n \leqslant y) \leqslant 2\exp\Big(-\frac{x^2}{2y}\Big). \tag{2.1}$$

We shall improve inequality (2.1) by showing that for any positive x and y,

$$\mathbb{P}(|M_n| \geqslant x, [M]_n + <M>_n \leqslant y) \leqslant 2\exp\Big(-\frac{8x^2}{9y}\Big). \tag{2.2}$$

Moreover, it was proven by Delyon (2009) that for any positive x and y,

$$\mathbb{P}(|M_n| \geqslant x, [M]_n + 2\,<M>_n \leqslant y) \leqslant 2\exp\Big(-\frac{3x^2}{2y}\Big). \tag{2.3}$$

We will show that inequality (2.3) is a special case of a more general result involving a suitable weighted sum of $[M]_n$ and $<M>_n$. Furthermore, it was shown by De la Peña and Pang (2009) that for any positive x,

$$\mathbb{P}\Big(\frac{|M_n|}{\sqrt{[M]_n + <M>_n + \mathbb{E}[M_n^2]}} \geqslant x\sqrt{\frac{3}{2}}\Big) \leqslant \Big(\frac{2}{3}\Big)^{1/3} x^{-2/3}\exp\Big(-\frac{x^2}{2}\Big). \tag{2.4}$$

We shall improve inequality (2.4) by using of the tailor-made normalization

$$S_n(a) = [M]_n + c(a)<M>_n, \tag{2.5}$$

where for any $a > 1/8$,

$$c(a) = \frac{2(1 - 2a + 2\sqrt{a(a+1)})}{8a - 1}. \tag{2.6}$$

The novelty of our approach is that $S_n(a)$ is a suitable weighted sum $<M>_n$ and $[M]_n$. For small values of n, the behavior of $<M>_n$ may be totally different from that of $[M]_n$. Consequently, our approach provides interesting concentration inequalities in many situations where $<M>_n \neq [M]_n$. The chapter is organised as follows. Section 2.2 is devoted to our new concentration inequalities for self-normalized martingales which improve some previous results of Bercu and Touati: Bercu and Touati (2008), Delyon (2009) and De la Peña and Pang (2009). Section 2.3 deals with statistical applications on autoregressive process, internal diffusion-limited aggregation process, and online statistical learning. All technical proofs are postponed to Sections 2.4 and 2.5.

2.2 Main results

Our first result holds without any additional assumption on (M_n).

Theorem 2.1. *Let (M_n) be a locally square integrable real martingale. Then, as soon as $a > 1/8$, we have for any positive x and y,*

$$\mathbb{P}(|M_n| \geqslant x, S_n(a) \leqslant y) \leqslant 2\exp\Big(-\frac{x^2}{2ay}\Big), \tag{2.7}$$

where $S_n(a) = [M]_n + c(a) <M>_n$ and $c(a)$ is given by (2.6).

Remark 2.2. *The function c is positive, strictly convex and $c(a) \sim 1/2a$ as a tends to infinity. Special values are given in Table 1.*

a	9/55	4/21	9/40	25/96	1/3	9/16	49/72	4/5
$c(a)$	10	6	4	3	2	1	4/5	2/3

Table 1. Special values of the function $c(a)$

In the special case where $<M>_n = [M]_n$, $S_n(a)$ reduces to $S_n(a) = (1 + c(a)) < M >_n$ and the best choice of a is clearly the one that minimizes $aS_n(a) = a(1 + c(a)) <M>_n$, that is $a = 1/3$.

Remark 2.3. *On the one hand, $c(a) = 1$ if and only if $a = 9/16$. Replacing the value $a = 9/16$ into (2.7) immediately leads to (2.2) as $S_n(a) = [M]_n + <M>_n$. On the other hand, $c(a) = 2$ if and only if $a = 1/3$. Hence, in this special case, $S_n(a) = [M]_n + 2 <M>_n$ and we find again (2.3) by taking the value $a = 1/3$ into (2.7).*

Our second result for self-normalized martingales is as follows.

Theorem 2.4. *Let (M_n) be a locally square integrable real martingale. Then, as soon as $a > 1/8$, we have for any positive x and y,*

$$\mathbb{P}\left(\frac{|M_n|}{S_n(a)} \geqslant x, S_n(a) \geqslant y\right) \leqslant 2\exp\Big(-\frac{x^2 y}{2a}\Big) \tag{2.8}$$

where $S_n(a) = [M]_n + c(a) <M>_n$ and $c(a)$ is given by (2.6). Moreover, we have for any positive x,

$$\mathbb{P}\Big(\frac{|M_n|}{S_n(a)} \geqslant x\Big) \leqslant 2 \inf_{p>1} \left(\mathbb{E}\Big[\exp\Big(-\frac{(p-1)x^2 S_n(a)}{2a}\Big)\Big]\right)^{1/p}. \tag{2.9}$$

Remark 2.5. *In the case $a = 9/16$, we find from (2.8) and (2.9) that for any positive x and y,*

$$\mathbb{P}\left(\frac{|M_n|}{[M]_n + <M>_n} \geqslant x, [M]_n + <M>_n \geqslant y\right) \leqslant 2\exp\Big(-\frac{8x^2y}{9}\Big),$$

$$\mathbb{P}\Big(\frac{|M_n|}{[M]_n + <M>_n} \geqslant x\Big) \leqslant 2\inf_{p>1}\left(\mathbb{E}\Big[\exp\Big(-\frac{8(p-1)x^2}{9}\Big([M]_n + <M>_n\Big)\Big)\Big]\right)^{1/p}.$$

Similar concentration inequalities for self-normalized martingales can be obtained for $a = 1/3$. In addition, via the same lines as in the proof of Theorem 2.4, we find that for any positive x and y,

$$\mathbb{P}\left(\frac{|M_n|}{<M>_n} \geqslant x, c(a)<M>_n \geqslant [M]_n + y\right) \leqslant 2\exp\Big(-\frac{x^2y}{2ac^2(a)}\Big), \tag{2.10}$$

$$\mathbb{P}\Big(\frac{|M_n|}{<M>_n} \geqslant x, [M]_n \leqslant c(a)y<M>_n\Big) \leqslant 2\inf_{p>1}\left(\mathbb{E}\Big[\exp\Big(-\frac{(p-1)x^2<M>_n}{2ac(a)(1+y)}\Big)\Big]\right)^{1/p}. \tag{2.11}$$

Our third result deals with missing factors in exponential inequalities for self-normalized martingales with upper bounds independent of $[M]_n$ or $<M>_n$.

Theorem 2.6. *Let (M_n) be a locally square integrable real martingale. Assume that $\mathbb{E}[|M_n|^p] < \infty$ for some $p \geqslant 2$. Then, as soon as $a > 1/8$, we have for any positive x,*

$$\mathbb{P}\Big(\frac{|M_n|}{\sqrt{aS_n(a) + (\mathbb{E}[|M_n|^p])^{2/p}}} \geqslant \frac{x}{\sqrt{B_q}}\Big) \leqslant C_q x^{-B_q}\exp\Big(-\frac{x^2}{2}\Big) \tag{2.12}$$

where $q = p/(p-1)$ is the Hölder conjugate exponent of p,

$$B_q = \frac{q}{2q-1} \qquad \text{and} \qquad C_q = \left(\frac{q}{2q-1}\right)^{B_q/2}.$$

In particular, for $p = 2$, we have for any positive x,

$$\mathbb{P}\Big(\frac{|M_n|}{\sqrt{aS_n(a) + \mathbb{E}[M_n^2]}} \geqslant x\sqrt{\frac{3}{2}}\Big) \leqslant \left(\frac{2}{3}\right)^{1/3} x^{-2/3}\exp\Big(-\frac{x^2}{2}\Big). \tag{2.13}$$

Remark 2.7. *In the case $a = 9/16$, we deduce from (2.13) that for any positive x,*

$$\mathbb{P}\Big(\frac{|M_n|}{\sqrt{a([M]_n + <M>_n) + \mathbb{E}[M_n^2]}} \geqslant x\sqrt{\frac{3}{2}}\Big) \leqslant \left(\frac{2}{3}\right)^{1/3} x^{-2/3}\exp\Big(-\frac{x^2}{2}\Big).$$

Since $a < 1$, this inequality clearly leads to

$$\mathbb{P}\Big(\frac{|M_n|}{\sqrt{[M]_n + <M>_n + \mathbb{E}[M_n^2]}} \geqslant x\sqrt{\frac{3}{2}}\Big) \leqslant \Big(\frac{2}{3}\Big)^{1/3} x^{-2/3} \exp\Big(-\frac{x^2}{2}\Big).$$

Consequently, in the case $a = 9/16$, (2.13) provides a tighter upper bound than inequality (2.4). Moreover, if $a = 1/3$, we obtain from (2.13) that for any positive x,

$$\mathbb{P}\Big(\frac{|M_n|}{\sqrt{[M]_n + 2\,<M>_n + 3\mathbb{E}[M_n^2]}} \geqslant \frac{x}{\sqrt{2}}\Big) \leqslant \Big(\frac{2}{3}\Big)^{1/3} x^{-2/3} \exp\Big(-\frac{x^2}{2}\Big).$$

Proof. The proofs are given in Sections 2.4 and 2.5. □

2.3 Statistical applications

2.3.1 Autoregressive process

Consider the first-order autoregressive process given, for all $n \geqslant 1$, by

$$X_n = \theta X_{n-1} + \varepsilon_n \tag{2.14}$$

where X_n and ε_n are the observation and the driven noise of the process, respectively. Assume that (ε_n) is a sequence of independent random variables sharing the same $\mathcal{N}(0, \sigma^2)$ distribution where $\sigma^2 > 0$. The process is said to be stable if $|\theta| < 1$, unstable if $|\theta| = 1$, and explosive if $|\theta| > 1$. We estimate the unknown parameter θ by the standard least-squares estimator given, for all $n \geqslant 1$, by

$$\widehat{\theta}_n = \frac{\sum_{k=1}^n X_{k-1} X_k}{\sum_{k=1}^n X_{k-1}^2}. \tag{2.15}$$

It is well-known that whatever the value of θ is, $\widehat{\theta}_n$ converges almost surely to θ. Moreover, White (1958) has shown that in the stable case $|\theta| < 1$,

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1 - \theta^2),$$

while in the explosive case $|\theta| > 1$ with initial value $X_0 = 0$,

$$|\theta|^n\Big(\widehat{\theta}_n - \theta\Big) \xrightarrow{\mathcal{L}} (\theta^2 - 1)\mathcal{C}$$

where $\mathcal{C}$ stands for the Cauchy distribution. Furthermore, in the stable case $|\theta| < 1$, it was proven in Bercu et al. (1997) that the sequence $(\widehat{\theta}_n)$ satisfies a large deviation principle with a convex-concave rate function. A fairly simple concentration inequality for the estimator $\widehat{\theta}_n$ was established in Bercu and Touati

(2008), whatever the value of θ is. More precisely, assume that X_0 is independent of (ε_n) with $\mathcal{N}(0, \tau^2)$ distribution where $\tau^2 \geqslant \sigma^2$. Then, for all $n \geqslant 1$ and for any positive x, we have

$$\mathbb{P}(|\widehat{\theta}_n - \theta| \geqslant x) \leqslant 2\exp\Big(-\frac{nx^2}{2(1+y_x)}\Big) \tag{2.16}$$

where y_x is the unique positive solution of the equation $h(y_x) = x^2$ and h is the function given, for any positive x, by $h(x) = (1+x)\log(1+x) - x$. It follows from (2.16) that, as soon as $0 < x < 1/2$,

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2\exp\Big(-\frac{nx^2}{2(1+2x)}\Big).$$

The situation in which (ε_n) is not normally distributed, is much more difficult to handle. If (ε_n) is a sequence of independent and identically distributed random variables, uniformly bounded with symmetric distribution, we can use De la Peña (1999)'s inequality for self-normalized conditionally symmetric martingales, to prove concentration inequalities for the least-squares estimator, see Bercu et al. (2015). Our motivation is to establish concentration inequalities for $\widehat{\theta}_n$ in the situation where the distribution of (ε_n) is non-symmetric.

Corollary 2.8. *Assume that (ε_n) is a sequence of independent and identically distributed random variables such that, for all $n \geqslant 1$,*

$$\varepsilon_n = \begin{cases} 2q & \text{with probability} \quad p, \\ -2p & \text{with probability} \quad q, \end{cases}$$

where $p \in]0, 1/2]$ and $q = 1-p$. Moreover, assume that X_0 is independent of (ε_n) with $|X_0| \geqslant 2p$. Then, for any $a > 1/8$ and for any x in the interval $[0, \sqrt{ad(a)}]$, we have

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2\exp\Big(-\frac{np^2x^2}{ad(a)}\Big) \qquad \text{where} \qquad d(a) = \frac{4\Big(q^2 + pqc(a)\Big)^2}{\Big(p^2 + pqc(a)\Big)}. \tag{2.17}$$

Remark 2.9. *In the symmetric case $p = 1/2$, we clearly have from (2.19) that $<M>_n = [M]_n$, $S_n(a) = (1 + c(a)) <M>_n$ and $d(a)$ reduces to $d(a) = 1 + c(a)$. Hence, if $a = 1/3$, $c(a) = 2$ and $d(a) = 3$. Consequently, we deduce from (2.17) that for any x in $[0, 1]$,*

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2\exp\Big(-\frac{nx^2}{4}\Big).$$

Moreover, in the nonsymmetric case $p \neq 1/2$, we always have $<M>_n \neq [M]_n$. For example, if $p = 1/3$ and $a = 9/16$, $c(a) = 1$ and $d(a) = 16/3$ which implies that $ad(a) = 3$. Therefore, we obtain from (2.17) that for any x in $[0, \sqrt{3}]$,

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2\exp\Big(-\frac{nx^2}{27}\Big).$$

Proof. It immediately follows from (2.14) together with (2.15) that for all $n \geqslant 1$,

$$\widehat{\theta}_n - \theta = \sigma^2 \frac{M_n}{<M>_n} \tag{2.18}$$

where $\sigma^2 = 4pq$ and (M_n) is the locally square integrable real martingale given by

$$M_n = \sum_{k=1}^{n} X_{k-1}\varepsilon_k, \qquad <M>_n = \sigma^2 \sum_{k=1}^{n} X_{k-1}^2, \qquad [M]_n = \sum_{k=1}^{n} X_{k-1}^2 \varepsilon_k^2. \tag{2.19}$$

We clearly have $\Big(c(a) + r\Big) <M>_n \leqslant S_n(a) \leqslant \Big(c(a) + r^{-1}\Big) <M>_n$ with $r = p/q$. Hence, we obtain from (2.9) that for any $a > 1/8$ and for any positive x,

$$\mathbb{P}\Big(|M_n| \geqslant x S_n(a)\Big) \leqslant 2 \left(\mathbb{E}\Big[\exp\Big(-\frac{x^2 S_n(a)}{2a}\Big)\Big]\right)^{1/2} \tag{2.20}$$

which implies via (2.18) that

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2 \left(\mathbb{E}\Big[\exp\Big(-\frac{x^2 <M>_n}{2a\sigma^2 d(a)}\Big)\Big]\right)^{1/2} \tag{2.21}$$

where $d(a)$ is given by (2.17). It only remains to find a suitable upper-bound for the Laplace transform of $<M>_n$. We have from (2.14) that $X_n^2 = \theta^2 X_{n-1}^2 + 2\theta X_{n-1}\varepsilon_n + \varepsilon_n^2$. Hence, if $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$, we obtain that for any real t and for all $n \geqslant 1$,

$$\mathbb{E}[\exp(tX_n^2)|\mathcal{F}_{n-1}] = \exp(t\theta^2 X_{n-1}^2)\Lambda_{n-1}(t) \tag{2.22}$$

where

$$\Lambda_{n-1}(t) = p\exp\Big(4tq^2 + 4\theta tqX_{n-1}\Big) + q\exp\Big(4tp^2 - 4\theta tpX_{n-1}\Big). \tag{2.23}$$

It follows from the so-called Kearns-Saul's inequality given in Lemma 2.36, page 37 of Bercu et al. (2015) that for any real s,

$$p\exp(qs) + q\exp(-ps) \leqslant \exp\Big(\frac{\varphi(p)s^2}{4}\Big), \tag{2.24}$$

where $\varphi(p) = (q-p)/\log(q/p) \in [0, 1/2]$. Then, we deduce from (2.23) and (2.24) with $s = 4\theta t X_{n-1}$ that for any $t \leqslant 0$, $\Lambda_{n-1}(t) \leqslant \exp\Big(4tp^2 + 4\varphi(p)t^2\theta^2 X_{n-1}^2\Big)$ leading to

$$\mathbb{E}[\exp(tX_n^2)|\mathcal{F}_{n-1}] \leqslant \exp\Big(4tp^2 + t\theta^2 X_{n-1}^2(1 + 4\varphi(p)t)\Big). \tag{2.25}$$

As soon as $t \in [-1/2, 0]$, we get from (2.25) that $\mathbb{E}[\exp(tX_n^2)|\mathcal{F}_{n-1}] \leqslant \exp(4tp^2)$. Consequently, for any $t \in [-1/2\sigma^2, 0]$ and for all $n \geqslant 1$,

$$\mathbb{E}[\exp(t <M>_n)] \leqslant \mathbb{E}[\exp(t <M>_{n-1})] \exp(4tp^2\sigma^2) \leqslant \exp(4ntp^2\sigma^2) \tag{2.26}$$

as $|X_0| \geqslant 2p$. Therefore, it follows from (2.21) and (2.26) that for any $x \in [0, \sqrt{ad(a)}]$,

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2\exp\Big(-\frac{np^2x^2}{ad(a)}\Big)$$

which achieves the proof of Corollary 2.8. □

Remark 2.10. *In the nonsymmetric case $p \neq 1/2$, we will set up some numerical simulations to highlight the theoretical results presented above.*

At first we plot the path of the process. Naturally, the process changes depending on the parameters, especially on θ. A p fixed, We simulate the cases: stable, unstable and explosive we can see it with the following plots:

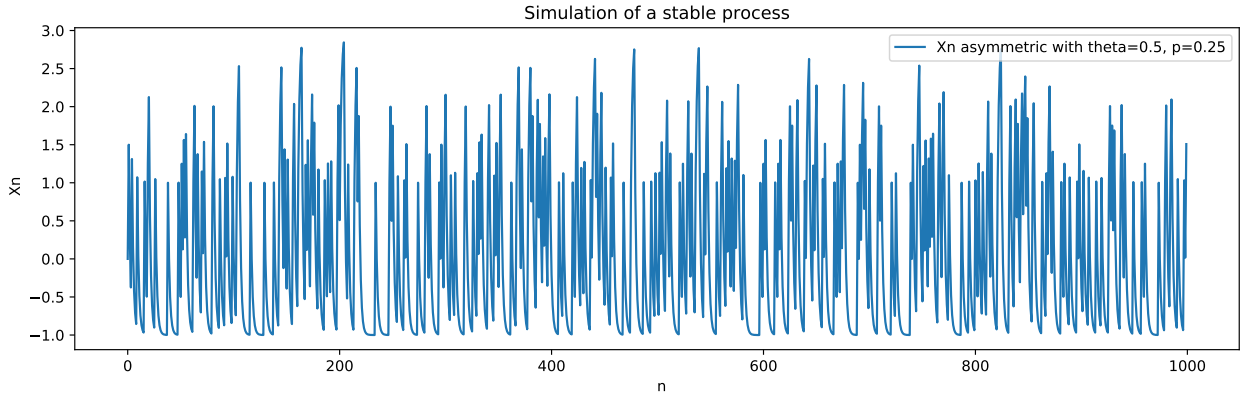


Figure 2.1 – Simulation of (X_n) $(p = 0.25,\ \theta = 0.5)$

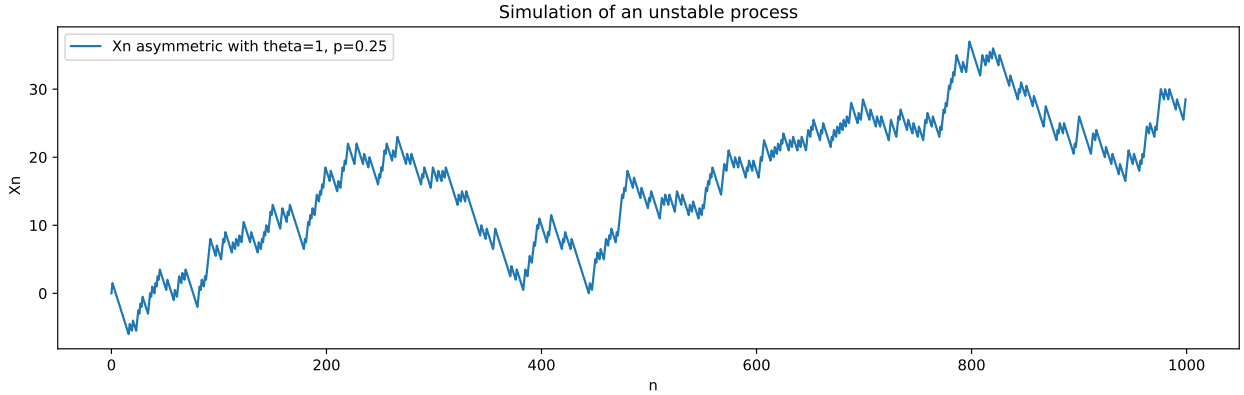


Figure 2.2 – Simulation of (X_n) $(p = 0.25,\ \theta = 1)$

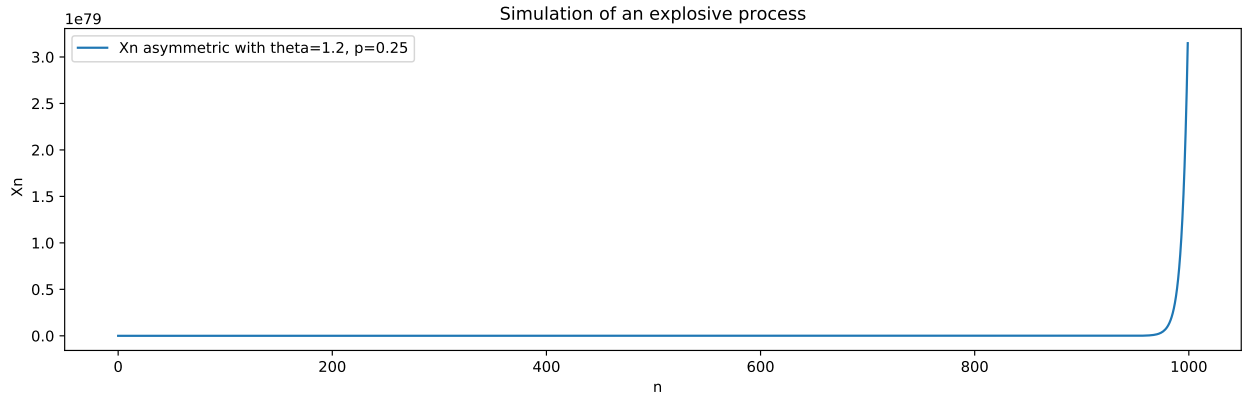


Figure 2.3 – Simulation of (X_n) $(p = 0.25,\ \theta = 1.2)$

We notice that the behavior of the asymmetric process is analogous to that of the symmetric process insofar as we find the 3 cases: stable, unstable and explosive.

Finally for the same θ we compare in the same plot a symmetric process with Gaussian error and the asymmetric process (Stable and unstable case)

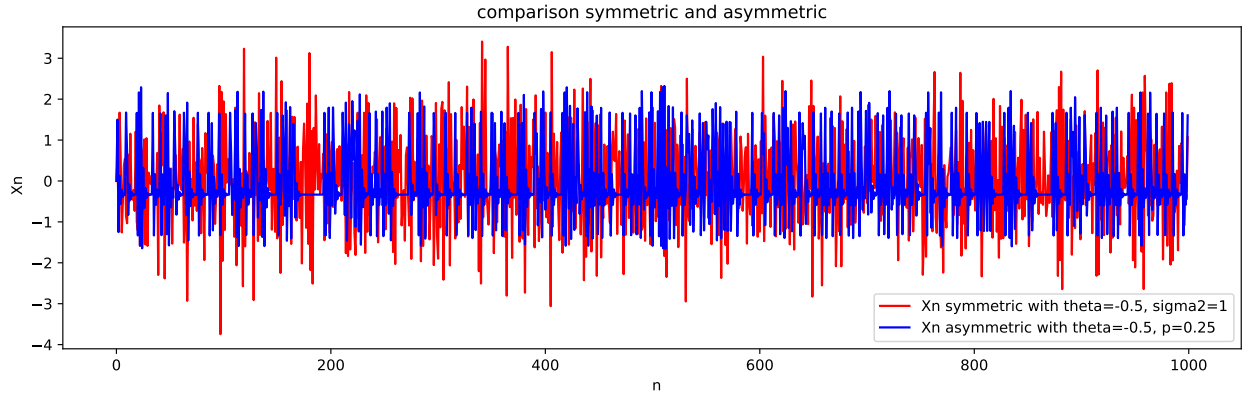


Figure 2.4 – Comparison of symmetric and assymetric process (stable case)

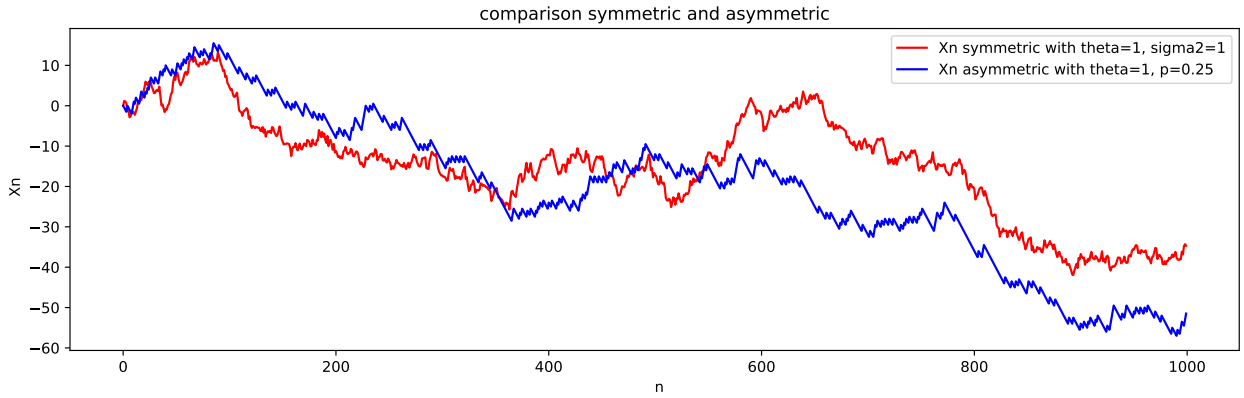


Figure 2.5 – Comparison of symmetric and assymetric process (unstable case)

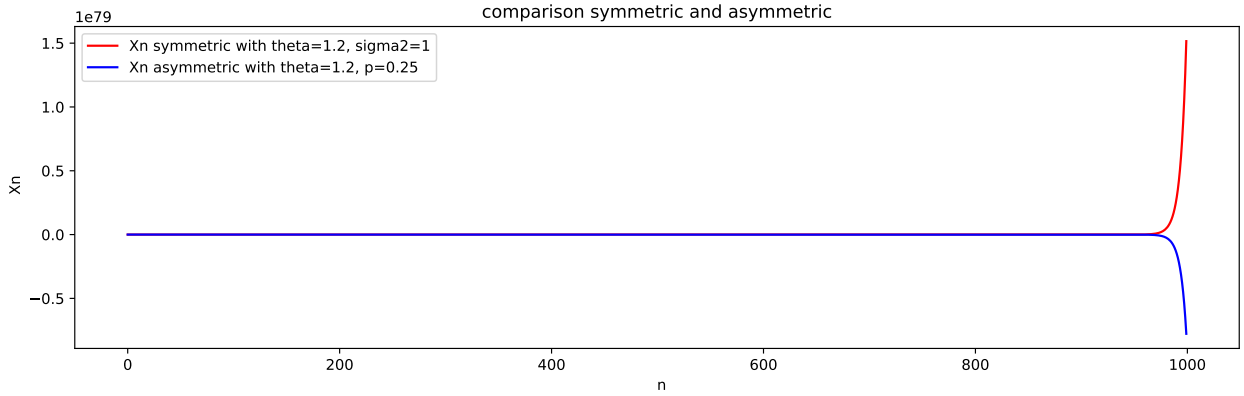


Figure 2.6 – Comparison of symmetric and assymetric process (explosive case)

These last three graphs show that although by keeping the three main properties i.e stable; unstable and explosive, the two processes remain fundamentally different.
Technically, it is more difficult to control the deviation of the least squares estimator in the case of a non-symmetric autoregressive process and it turns out that our results are more suitable and more efficient.
We will highlight this at the end of the section through explicit calculations and an objective comparison with the inequalities in the literature.

We illustrate the almost sure convergence for different values of θ.

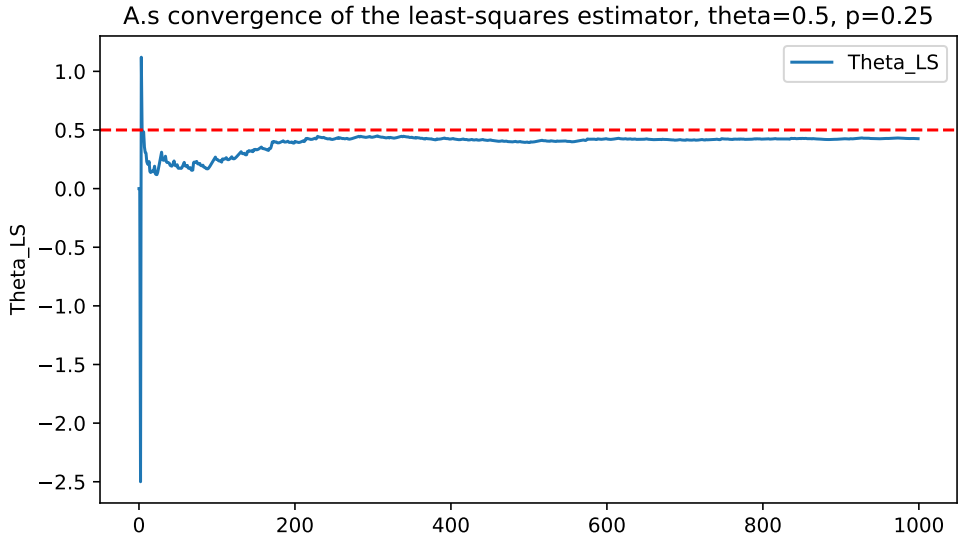


Figure 2.7 – Almost surely convergence of $\widehat{\theta}_n$ (Stable case)

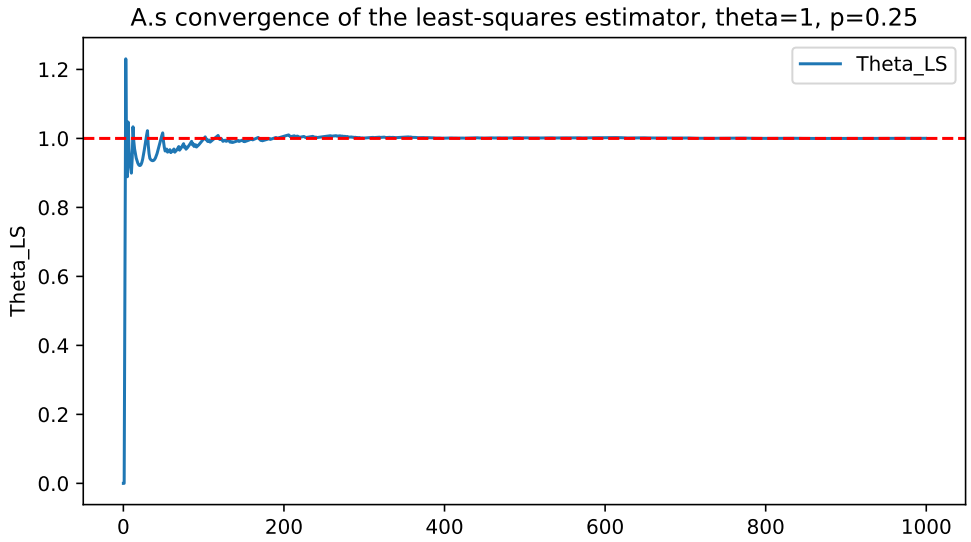


Figure 2.8 – Almost surely convergence of $\widehat{\theta}_n$ (Unstable case)

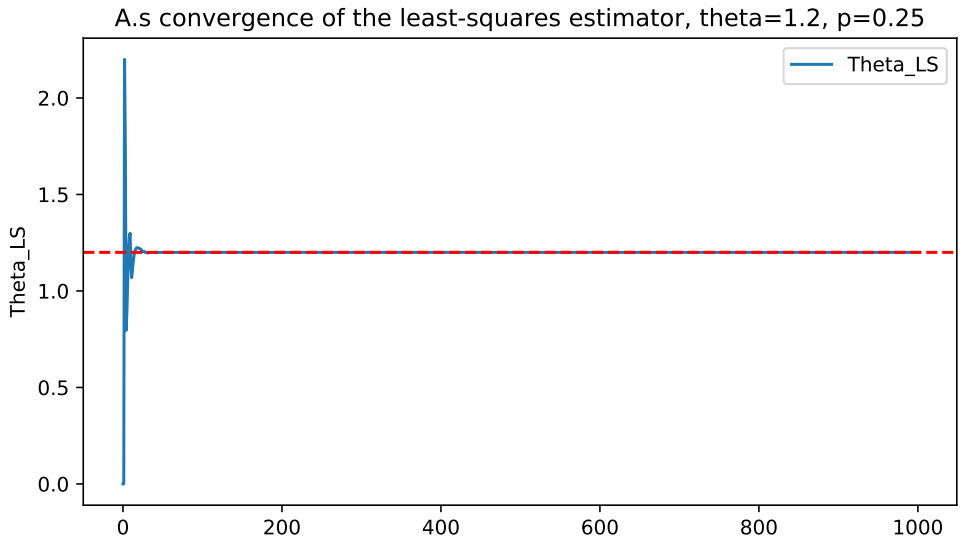


Figure 2.9 – Almost surely convergence of $\widehat{\theta}_n$ (Explosive case)

One of the major properties of this process is the a.s convergence of the estimator with respect to the theoretical value in the stable, unstable and explosive case. This is well highlighted in the three plots above, in which we can visually appreciate this convergence.

We recall that according to White (1958), in the stable case $|\theta| < 1$,

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1 - \theta^2).$$

The next two plots highlight the asymptotic normality of $\widehat{\theta}_n$.

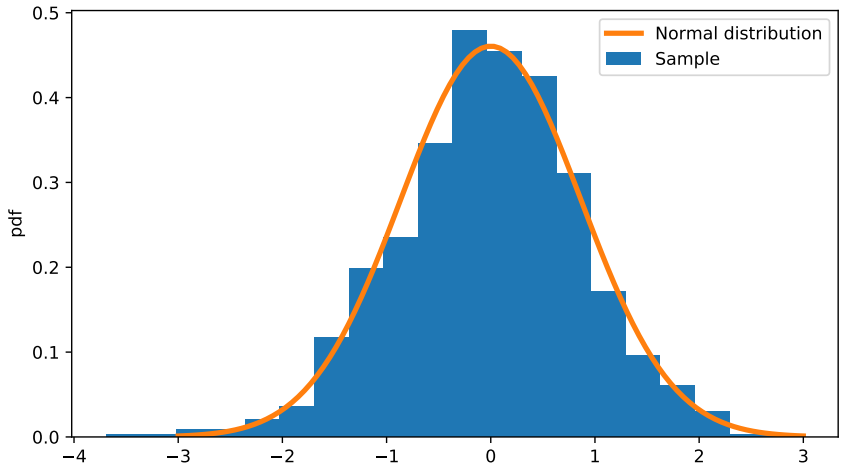


Figure 2.10 – Asymptotic normality of $\widehat{\theta}_n$ ($p = 0.3$, $\theta = 0.5$)

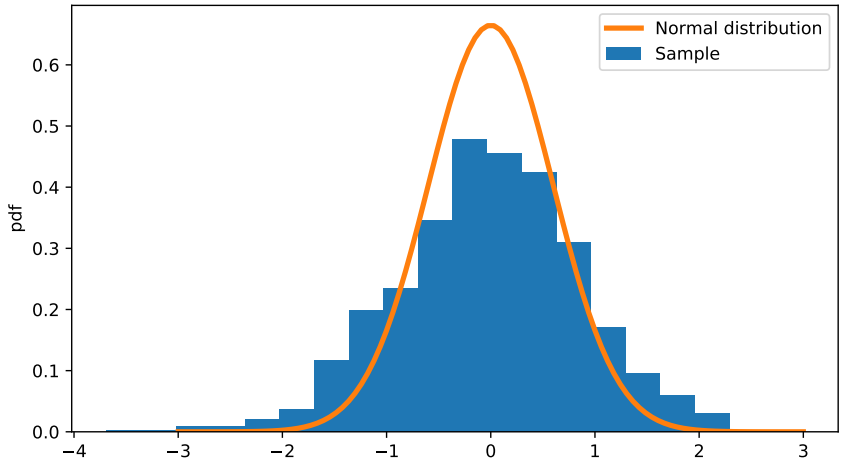


Figure 2.11 – Asymptotic normality of $\widehat{\theta}_n$ ($p = 0.3$, $\theta = 0.8$)

We see explicitly via these plots the impact of the parameter θ on the variance, namely the smaller the parameter the greater the variance. If we get closer to the unstable case, the variance is in a neighborhood of zero.

We obtain from (2.17) ($a = \frac{9}{16}$ and $\frac{1}{3}$ respectively) that for any x in $[0, \sqrt{3}]$.

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2\exp\Big(-\frac{nx^2}{27}\Big).$$

$$\mathbb{P}\Big(|\widehat{\theta}_n - \theta| \geqslant x\Big) \leqslant 2\exp\Big(-\frac{15n.x^2}{2304}\Big).$$

At p fixed, we can always find the optimal bound, this amounts to minimizing a*d(a). Numerically we get for, $p = \frac{1}{3}$ $a^* = 0.2216034, d(a^*) = 7.217284$. The following plot illustrates the different bounds:

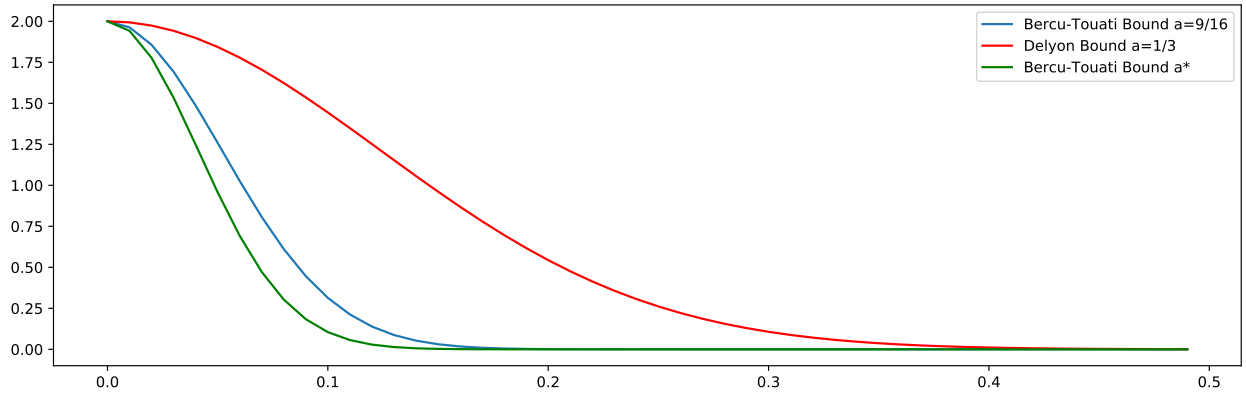


Figure 2.12 – Deviation bounds for $|\widehat{\theta}_n - \theta|$

This illustrates the importance of the results 2.1. For this process, we have tighter inequalities than that obtained via Delyon (2009). Another advantage of our parametrization (dependent of a) is that we can identify an optimal inequality.
This type of bounds is quite technical to set up but allows control over $|\widehat{\theta}_n - \theta|$ for a more complicated autoregressive process than that studied by Bercu and Touati (2008). We are thus able to strengthen the tools previously mentioned and provide them with more flexibility to adapt them to more applications.

2.3.2 Internal diffusion-limited aggregation process

Our second application deals with the internal diffusion-limited aggregation process. This aggregation process, first introduced in Mathematics by Diaconis and Fulton Diaconis and Fulton (1991), is a cluster growth model in $\mathbb{Z}^d$ where explorers, starting from the origin at time 0, are travelling as a simple random walk on $\mathbb{Z}^d$ until they reach an uninhabited site that is added to the cluster. In the special case $d = 1$, the cluster is an interval $A(n) = [L_n, R_n]$ which, properly normalized, converges almost surely to $[-1, 1]$. In dimension $d \geqslant 2$, Lawler,

Bramson and Griffeath Lawler et al. (1992) have shown that the limit shape of the cluster is a sphere. We shall restrict our attention on the one-dimensional internal diffusion-limited aggregation process. Consider the simple random walk on the integer number line $\mathbb{Z}$ starting from the origin at time 0.

At each step, the explorer moves to the right $+1$ or to the left -1 with equal probability $1/2$. Let $(A(n))$ be the sequence of random subsets of $\mathbb{Z}$, recursively defined as follows: $A(0) = \{0\}$ and, for all $n \geqslant 0$,

$$A(n+1) = \begin{cases} A(n) \cup \{L_n - 1\} \\ A(n) \cup \{R_n + 1\} \end{cases}$$

if the explorer leaves $A(n)$ by the left side or by the right side, respectively, where L_n and R_n stand for being the minimum and the maximum of $A(n) = \{L_n, L_n + 1, \ldots, R_n - 1, R_n\}$. The random set $A(n)$ is characterized by $X_n = L_n + R_n$ as $R_n - L_n = n$. One can observe that L_n and R_n correspond to the number of negative and positive sites of $A(n)$, respectively. It was proven in Diaconis and Fulton (1991) that

$$\lim_{n\to\infty} \frac{X_n}{n} = 0 \qquad \text{a.s.}$$

and

$$\frac{X_n}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{1}{3}\Big).$$

It is possible to prove from Azuma-Hoeffding's inequality Bercu et al. (2015) that for any positive x,

$$\mathbb{P}\Big(\frac{|X_n|}{n} \geqslant x\Big) \leqslant 2\exp\Big(-\frac{3}{8}nx^2\Big). \tag{2.27}$$

Our goal is to improve this inequality with a suitable use of Theorems 2.1 and 2.6.

Corollary 2.11. *For any a in the interval $]1/8, 9/16]$ and for any positive x, we have*

$$\mathbb{P}\Big(\frac{|X_n|}{n} \geqslant x\Big) \leqslant 2\exp\Big(-\frac{nx^2}{2ac_n(a)}\Big) \tag{2.28}$$

and

$$\mathbb{P}\Big(\frac{|X_n|}{\sqrt{n}} \geqslant x\Big) \leqslant (d_n(a))^{1/3} x^{-2/3} \exp\Big(-\frac{x^2}{3d_n(a)}\Big) \tag{2.29}$$

where

$$c_n(a) = \Big(\frac{2n+1}{n+1}\Big)\Big(\frac{3+c(a)}{6}\Big) + \Big(\frac{n(1+c(a)) + 2c(a)}{(n+1)^2}\Big), \; d_n(a) = c_n(a) + \Big(\frac{n+2}{3n}\Big). \tag{2.30}$$

Remark 2.12. *The calculation of $c_n(a)$ and $d_n(a)$ is quite straightforward. As a matter of fact, if $a = 1/3$, $c(a) = 2$ and it immediately follows from (2.30) that for all $n \geqslant 1$ $c_n(a) \leqslant 3$ and $d_n(a) \leqslant 4$. We can deduce from (2.28) that for any positive x,*

$$\mathbb{P}\Big(\frac{|X_n|}{n} \geqslant x\Big) \leqslant 2\exp\Big(-\frac{nx^2}{2}\Big)$$

which clearly outperforms inequality (2.27). In addition, (2.29) implies that for any positive x,

$$\mathbb{P}\Big(\frac{|X_n|}{\sqrt{n}} \geqslant x\Big) \leqslant \Big(\frac{2}{x}\Big)^{2/3}\exp\Big(-\frac{x^2}{12}\Big).$$

Moreover, if $a = 25/96$, $c(a) = 3$ and we obtain from (2.30) that for all $n \geqslant 1$, $c_n(a) \leqslant 7/2$ and $d_n(a) \leqslant 9/2$. We find from (2.28) that for any positive x,

$$\mathbb{P}\Big(\frac{|X_n|}{n} \geqslant x\Big) \leqslant 2\exp\Big(-\frac{96nx^2}{175}\Big).$$

It improves the above inequality for $a = 1/3$. Finally, we deduce from (2.29) that for any positive x,

$$\mathbb{P}\Big(\frac{|X_n|}{\sqrt{n}} \geqslant x\Big) \leqslant \Big(\frac{3}{\sqrt{2}x}\Big)^{2/3}\exp\Big(-\frac{2x^2}{27}\Big).$$

Proof. It follows from a stopping time argument for gambler's ruin that for all $n \geqslant 1$, $X_n = X_{n-1} + \xi_n$ where the distribution of ξ_n given $\mathcal{F}_{n-1}$ is a Rademacher $\mathcal{R}(p_n)$ distribution with

$$p_n = \frac{(n + 1 - X_{n-1})}{2(n+1)}.$$

Hence, we clearly have

$$\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1} + \mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = \Big(\frac{n}{n+1}\Big)X_{n-1} \tag{2.31}$$

and

$$\mathbb{E}[X_n^2|\mathcal{F}_{n-1}] = X_{n-1}^2 + 2X_{n-1}\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] + 1 = 1 + \Big(\frac{n-1}{n+1}\Big)X_{n-1}^2. \tag{2.32}$$

Let (M_n) be the sequence defined by $M_n = (n+1)X_n$. We immediately deduce from (2.31) and (2.32) that (M_n) is a locally square integrable real martingale such that

$$<M>_n = \sum_{k=1}^{n}(k+1)^2 - \sum_{k=1}^{n}X_{k-1}^2.$$

Moreover, for all $n \geqslant 1$, $|X_n| \leqslant n$. Hence,

$$[M]_n = \sum_{k=1}^{n}((k+1)X_k - kX_{k-1})^2 = \sum_{k=1}^{n}(k\xi_k + X_k)^2 \leqslant 3\sum_{k=1}^{n}k^2 + \sum_{k=1}^{n}X_k^2.$$

One can observe that we always have $<M>_n \neq [M]_n$. In addition,

$$S_n(a) \leqslant (3+c(a))\sum_{k=1}^{n} k^2 + (1-c(a))\sum_{k=1}^{n-1} X_k^2 + X_n^2 + c(a)n(n+2). \tag{2.33}$$

For any $a \in]1/8, 9/16]$, $c(a) \geqslant 1$. Therefore, we obtain from (2.33) that for any $a \in]1/8, 9/16]$,

$$S_n(a) \leqslant (3+c(a))\sum_{k=1}^{n} k^2 + n(n+c(a)(n+2)) \leqslant n(n+1)^2 c_n(a) \tag{2.34}$$

where $c_n(a)$ is given by (2.30). Hence, it follows from (2.7) with $y = n(n+1)^2 c_n(a)$ that for any $a \in]1/8, 9/16]$ and for any positive x,

$$\mathbb{P}\Big(\frac{|X_n|}{n} \geqslant x\Big) = \mathbb{P}\Big(|M_n| \geqslant xn(n+1), S_n(a) \leqslant y\Big) \leqslant 2\exp\Big(-\frac{nx^2}{2ac_n(a)}\Big),$$

which is exactly inequality (2.28). Furthermore, we can deduce from identity (2.32) that for all $n \geqslant 1$,

$$\mathbb{E}[X_n^2] = \frac{(n+2)}{3} \qquad \text{and} \qquad \mathbb{E}[M_n^2] = \frac{(n+1)^2(n+2)}{3}. \tag{2.35}$$

Finally, we find from (2.13) together with (2.30), (2.34) and (2.35) that for any $a \in]1/8, 9/16]$ and for any positive x,

$$\mathbb{P}\Big(\frac{|X_n|}{\sqrt{nd_n(a)}} \geqslant x\sqrt{\frac{3}{2}}\Big) \leqslant \Big(\frac{2}{3}\Big)^{1/3} x^{-2/3}\exp\Big(-\frac{x^2}{2}\Big)$$

which clearly leads to (2.29), completing the proof of Corollary 2.11. □

Remark 2.13. *In dimension 1, we recall that the diffusion-limited aggregation process have some asymptotic properties.*

$$\lim_{n\to\infty} \frac{X_n}{n} = 0 \; a.s$$

$$\frac{X_n}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{3}\right)$$

The following plots illustrate the trajectory of the process,its almost surely convergence and its asymptotic normality.

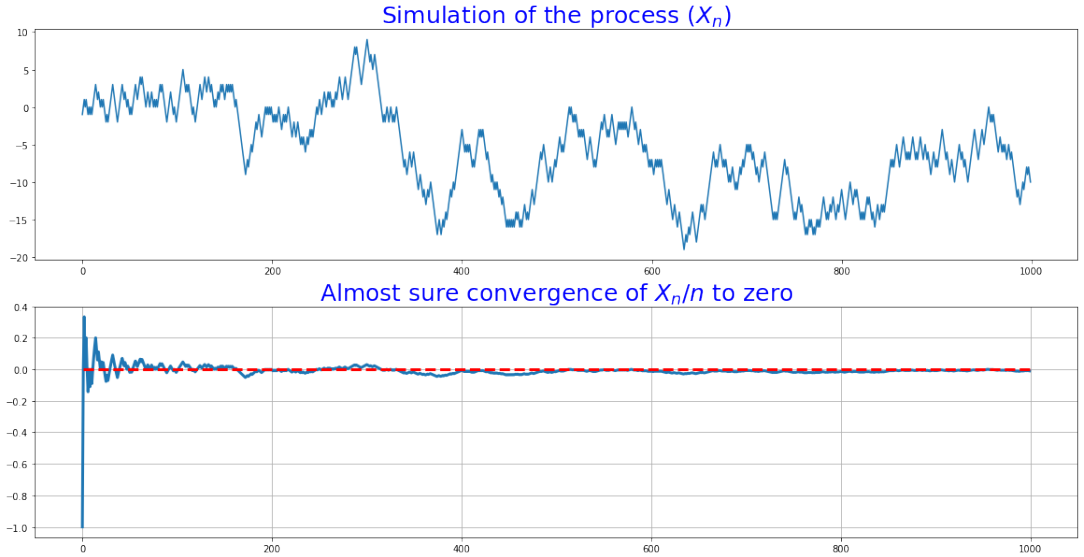


Figure 2.13 – Trajectory/ A.S convergence of a diffusion-limited aggregation process

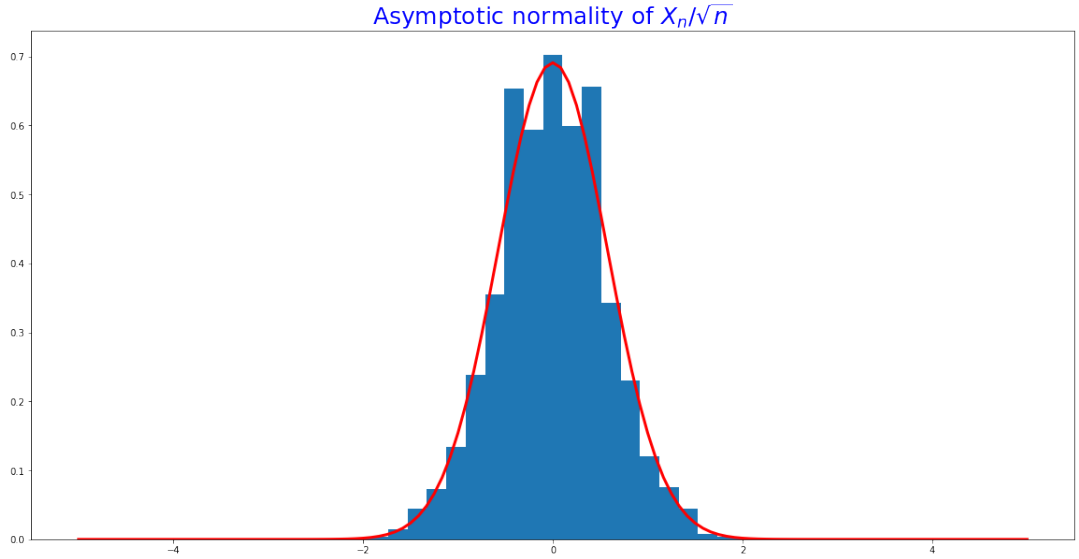


Figure 2.14 – Asymptotic normality of a diffusion-limited aggregation process

Referring to the remark 1.12, we have the following inequalities:
from 2.28 we have:

$$\mathrm{P}\left(\frac{|X_n|}{n} \geqslant x\right) \leqslant 2\exp\left(-\frac{3}{8}nx^2\right)$$

(By using Azuma inequality)

$$\mathrm{P}\left(\frac{|X_n|}{n} \geqslant x\right) \leqslant 2\exp\left(-\frac{nx^2}{2}\right)$$

(By using Delyon inequality $a = \frac{1}{3}$)

$$\mathrm{P}\left(\frac{|X_n|}{n} \geqslant x\right) \leqslant 2\exp\left(-\frac{96.nx^2}{175}\right)$$

(By using Bercu-Touati inequality $a = \frac{25}{96}$)
From 2.29 we have:

$$\mathrm{P}\left(\frac{|X_n|}{\sqrt{n}} \geqslant x\right) \leqslant \left(\frac{2}{x}\right)^{2/3}\exp\left(-\frac{x^2}{12}\right)$$

(By using Bercu-Touati inequality $a = \frac{1}{3}$)

$$\mathrm{P}\left(\frac{|X_n|}{\sqrt{n}} \geqslant x\right) \leqslant \left(\frac{3}{\sqrt{2}x}\right)^{2/3}\exp\left(-\frac{2x^2}{27}\right)$$

We illustrate these bounds through the following two plots.

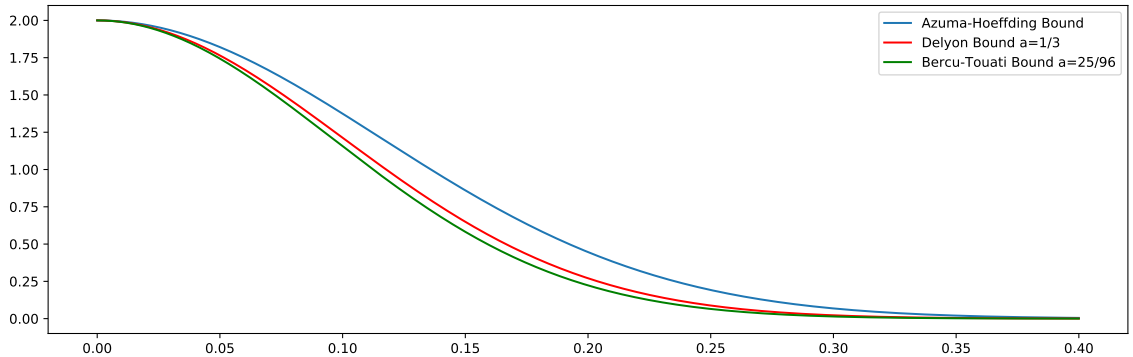


Figure 2.15 – Deviation bounds for $\frac{X_n}{n}$

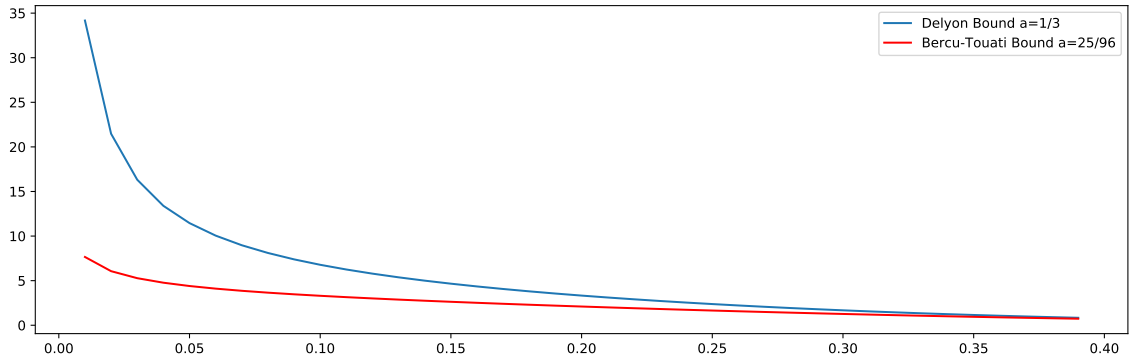


Figure 2.16 – Deviation bounds for $\frac{X_n}{\sqrt{n}}$

This confirms that we can obtain tighter bounds than that resulting from the inequality of Delyon (2009).
This is one more argument in favor of our new inequalities which, for each statistical application, provide tighter bounds, flexibility and the possibility of obtaining in each case an optimal bound.

2.3.3 Online statistical learning

Our third application is devoted to the study of the statistical risk of hypothesis during an online learning process using concentration inequalities for martingales. We refer the reader to the survey of Cesa-Bianchi and Lugosi (2006) for a rather exhaustive description of the underlying theory concerning online learning. Our approach is based on the contributions of Cesa-Bianchi et al. (2004), Cesa-Bianchi and Gentile (2008) dealing with the statistical risk of hypothesis in the situation where the ensemble of hypotheses is produced by training a learning algorithm incrementally on a data set of independent and identically distributed random variables. Their bounds rely on Freedman concentration inequality for martingales Freedman (1975). Consider the task of predicting a sequence in an online manner with inputs and outputs taking values in some abstract measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. We call hypothesis H, the classifier or regressor generated by a learning algorithm after training. The predictive performance of hypothesis H is evaluated by the theoretical risk denoted $R(H)$, which is the expected loss on a realisation $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ drawn from the underlying distribution

$$R(H) = \mathbb{E}[\ell(H(X), Y)]$$

where ℓ is a nonnegative and bounded loss function. For the sake of simplicity, we assume that ℓ is bounded by 1. Denote by $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ a training data set of independent random variables sharing the same unknown distribution as (X, Y). Our goal is to predict $Y_{n+1} \in \mathcal{Y}$ given $X_{n+1} \in \mathcal{X}$, on the

basis of $\mathcal{S}_n$. Let $\mathcal{H}_n = \{H_0, H_1, \ldots, H_{n-1}\}$ be a finite ensemble of hypotheses generated by an online learning algorithm where the initial hypothesis H_0 is arbitrarily chosen.The empirical risk and the average risk associated with the ensemble of hypotheses $\mathcal{H}_n$ and the training data set $\mathcal{S}_n$ are respectively given by

$$\widehat{R}_n = \frac{1}{n}\sum_{k=1}^{n} \ell(H_{k-1}(X_k), Y_k) \qquad \text{and} \qquad R_n = \frac{1}{n}\sum_{k=1}^{n} R(H_{k-1}). \tag{2.36}$$

Our bound on the average risk R_n is as follows.

Corollary 2.14. *Let $\mathcal{H}_n = \{H_0, H_1, ..., H_{n-1}\}$ be a finite ensemble of hypotheses generated by a learning algorithm. Then, for any a in the interval $]1/8, 9/16]$ and for any positive x, we have*

$$\mathbb{P}(R_n \geqslant \widehat{R}_n + x) \leqslant \exp\Big(-\frac{nx^2}{2a(1 + c(a)V_n)}\Big), \tag{2.37}$$

where

$$V_n = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[\ell^2(H_{k-1}(X), Y)]. \tag{2.38}$$

In other words, for any $0 < \delta \leqslant 1$,

$$\mathbb{P}\Big(R_n \geqslant \widehat{R}_n + \sqrt{\frac{2a(1 + c(a)V_n)\log(1/\delta)}{n}}\Big) \leqslant \delta. \tag{2.39}$$

Moreover, denote $m(a) = \max(4(1 + c(a)), c^2(a))/2$. Then, for any $0 < \delta \leqslant 1$ and for all integer $n \geqslant am(a)\log(1/\delta)$, we also have

$$\mathbb{P}\Big(R_n \geqslant \widehat{R}_n + \frac{ac(a)\log(1/\delta)}{n} + \sqrt{\frac{a\,\Delta_n(a)\log(1/\delta)}{n}}\Big) \leqslant \delta \tag{2.40}$$

where $\Delta_n(a) = 2 + 2c(a)\widehat{R}_n + ac^2(a)\log(1/\delta)/n$.

Remark 2.15. *On the one hand, (2.39) improves the deviation inequality given in Proposition 1 of Cesa-Bianchi et al. (2004),*

$$\mathbb{P}\Big(R_n \geqslant \widehat{R}_n + \sqrt{\frac{2\log(1/\delta)}{n}}\Big) \leqslant \delta,$$

as V_n is always smaller than 1. On the other hand, (2.40) is drastically more accurate than the deviation inequality given in Proposition 2 of Cesa-Bianchi and Gentile (2008),

$$\mathbb{P}\Big(R_n \geqslant \widehat{R}_n + \frac{36}{n}\log\Big(\frac{n\widehat{R}_n + 3}{\delta}\Big) + 2\sqrt{\frac{\widehat{R}_n}{n}\log\Big(\frac{n\widehat{R}_n + 3}{\delta}\Big)}\Big) \leqslant \delta. \tag{2.41}$$

Indeed, one can observe that the right-hand sides of (2.40) and (2.41) are increasing functions of $\widehat{R}_n$. The smallest value in (2.41) for $\widehat{R}_n = 0$ is given by $36\log(3/\delta)/n$. Consequently, inequality (2.41) is only effective for $n \geqslant 36\log(3/\delta)$, which implies that n must always be greater than 40. For example, if $\delta = 1/5$, it is necessary to assume that $n \geqslant 36\log(15)$, that is $n \geqslant 98$. If $a = 1/3$, then $c(a) = 2$ and $m(a) = 6$. Consequently, inequality (2.40) is interesting as soon as $n \geqslant -2\log(\delta)$. For example, if $\delta = 1/5$, it is necessary to assume that $n \geqslant 4$. For instance, if $\delta = 1/5$, $n = 100$ and $a = 1/3$, the smallest values in (2.40) and (2.41) are respectively given by 0.220 and 0.975. Finally, for all values of δ, n and a, one can easily check that (2.40) is always sharper than (2.41).

Proof. Let (M_n) be the locally square integrable real martingale given by

$$M_n = \sum_{k=1}^{n} \Big(R(H_{k-1}) - \ell(H_{k-1}(X_k), Y_k) \Big), \tag{2.42}$$

where we recall that $R(H) = \mathbb{E}[\ell(H(X), Y)]$. We clearly have

$$<M>_n = \sum_{k=1}^{n} \Big(\mathbb{E}[\ell^2(H_{k-1}(X), Y))] - R^2(H_{k-1}) \Big), \quad [M]_n = \sum_{k=1}^{n} \Big(R(H_{k-1}) - \ell(H_{k-1}(X_k), Y_k) \Big)^2.$$

Consequently, for any $a \in]1/8, 9/16]$,

$$S_n(a) \leqslant (1-c(a)) \sum_{k=1}^{n} R^2(H_{k-1}) + \sum_{k=1}^{n} \ell^2(H_{k-1}(X_k), Y_k) + c(a) \sum_{k=1}^{n} \mathbb{E}[\ell^2(H_{k-1}(X), Y)]$$

Hence, as $c(a) \geqslant 1$ and ℓ is bounded by 1, we obtain from (2.38) that $S_n(a) \leqslant n(1 + c(a)V_n)$. Therefore, it follows from (2.7) with $y = n(1 + c(a)V_n)$ that for any $a \in]1/8, 9/16]$ and for any positive x,

$$\mathbb{P}\Big(\frac{M_n}{n} \geqslant x \Big) \leqslant \exp\Big(-\frac{nx^2}{2a(1 + c(a)V_n)} \Big). \tag{2.43}$$

However, we clearly have from (2.42) that $M_n = n\Big(R_n - \widehat{R}_n\Big)$. Hence, (2.43) immediately implies (2.37) and (2.39). It only remains to prove (2.40). Since ℓ is bounded by 1, we obtain from (2.38) that $V_n \leqslant R_n$. Consequently, (2.39) ensures that for any $0 < \delta \leqslant 1$,

$$\mathbb{P}\Big(\Phi_a(R_n) \geqslant \widehat{R}_n \Big) \leqslant \delta \tag{2.44}$$

where the function Φ_a is defined, for all x in $[0, 1]$, by

$$\Phi_a(x) = x - \sqrt{\frac{2a(1 + c(a)x)\log(1/\delta)}{n}}.$$

It is not hard to see that, as soon as $n \geqslant am(a)\log(1/\delta)$ with $m(a) = \max(4(1+c(a)), c^2(a))/2$, Φ_a is a strictly convex and increasing function on $[0,1]$. Then, Φ_a is invertible and it follows from straightforward calculations that

$$\Phi_a^{-1}(x) = x + \frac{ac(a)\log(1/\delta)}{n} + \sqrt{\frac{a\log(1/\delta)}{n}\Big(2 + 2c(a)x + \frac{ac^2(a)\log(1/\delta)}{n}\Big)}.$$

Finally, we immediately obtain from (2.44) that

$$\mathbb{P}\Big(\Phi_a(R_n) \geqslant \widehat{R}_n\Big) = \mathbb{P}\Big(R_n \geqslant \Phi_a^{-1}(\widehat{R}_n)\Big) \leqslant \delta \tag{2.45}$$

which is exactly inequality (2.40), completing the proof of Corollary 2.14. □

Remark 2.16. *We use the previous corollary in order to derive a bound on the mean risk of the hypotheses generated by the Pegasos Shalev-Shwartz et al. (2011) algorithm. Pegasos is intrinsically linked to Support Vector Machines. It was designed to solve the underlying optimization problem.*

PEGASOS, Primal Estimated sub-GrAdient SOlver Shalev-Shwartz et al. (2011) is an online learning for support vector machine (SVM) scheme based on stochastic gradient decent (SGD) Bottou (2010).

Given the training set $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ the online learning PEGASOS algorithm aims at finding the minimizer of the primal SVM problem.

$$\min_{\omega \in \mathbb{R}^d} \left\{ \frac{1}{T} \sum_{t=1}^{T} \ell(y_t, h_t(x)) + \frac{\lambda}{2}\|\omega\|_2^2 \right\}$$

where $h_t(x) = \omega^\top \mathbf{x}$, $\ell(y, y') = \max(0, 1 - yy')$ is the hinge loss, and T the maximum training epoch. The pseudo code of an online algorithm with PEGASOS is summarized in Algorithm 1.

Experimental setting. In order to compare the upper bounds for the case of online learning algorithm proposed in this chapter with the base line upper bounds given in Cesa-Bianchi and Gentile (2008), we conduct numerical experiments on synthetic dataset. We set $n = 1000$ and for each $y_i \in \{-1, +1\}$ we draw samples

$x_i \sim \mathcal{N}((0,0)^\top, \left(\begin{smallmatrix} 1.1 & 0.1 \\ 0.1 & 1.1 \end{smallmatrix}\right))$ for $i \in \{1, \ldots, 500\}$ and $x_i \sim \mathcal{N}((2,2)^\top, \left(\begin{smallmatrix} 1.2 & 0.2 \\ 0.2 & 1.2 \end{smallmatrix}\right))$ for $i \in \{500, 1000\}$. In Figure 2.17, we plot objective function of PEGASOS algorithm during a 5-fold cross validation. We further display the accuracy performance for this binary classification problem where the regularization parameter is fixed as $\lambda = 10$. We notice that the data is linearly separable.

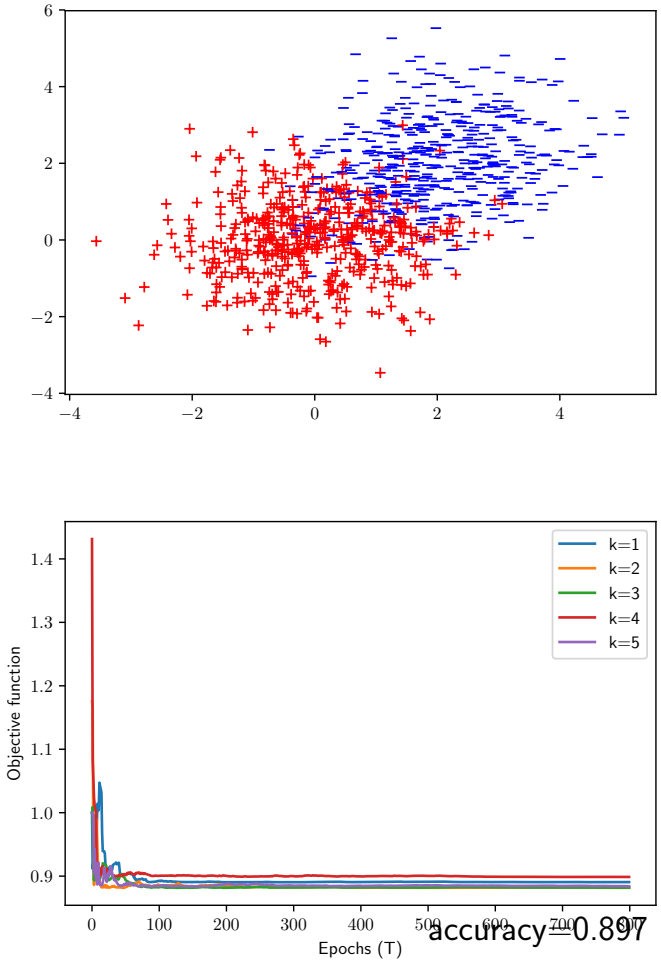


Figure 2.17 – (Top) Scatter plot of the synthetic data for binary classification according the describing setting above; (Bottom) Plots of the objective function of PEGASOS algorithm in a 5-fold cross validation. In the right bottom we display the accuracy for classification that corresponds to the mean of accuracies given by the 5-fold cross validation.

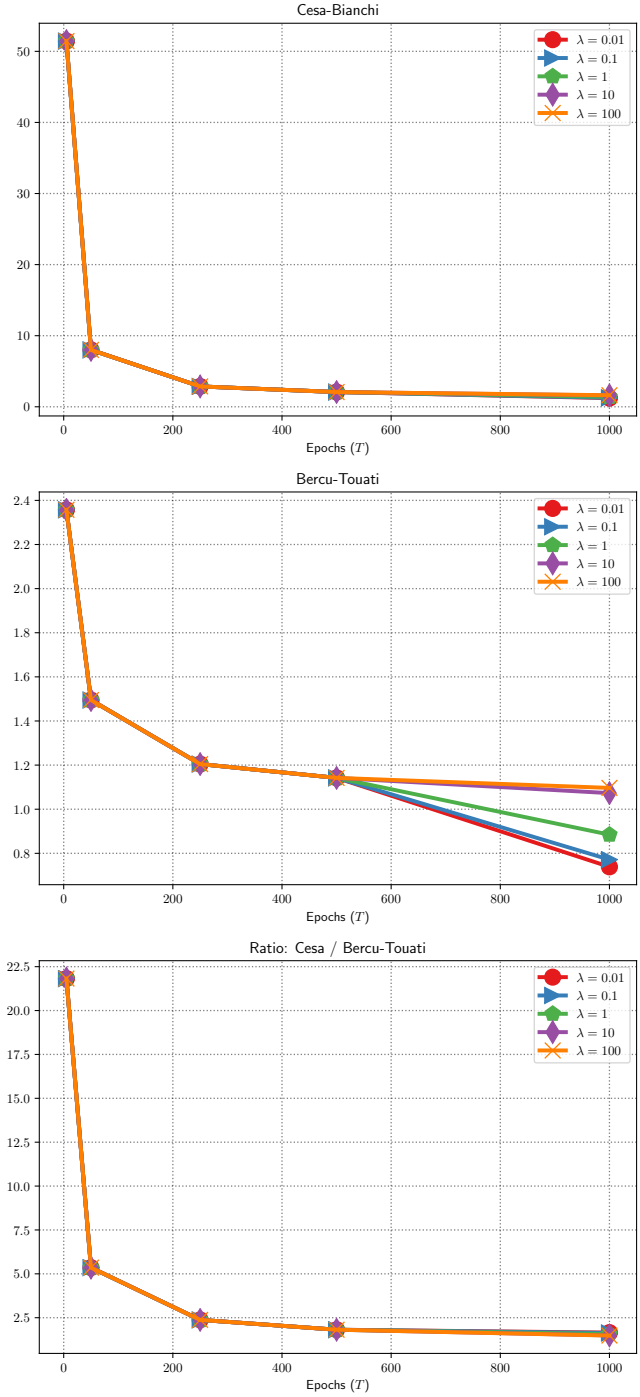


Figure 2.18 – Plots of the upper bounds on the average risk for online learning with PEGAOS algorithm as a function of the maximum numbers of epochs T (Top) for Cesa-Bianchi (Middle) for Bercu-Touati (Bottom) the ratio between Csea-Bianchi and Bercu-Touati upper bounds.

Algorithm 1 Online learning with PEGASOS

input: $h_0(\mathbf{x}) = \omega_0^\top \mathbf{x}$, where ω_0 a warm-start weighted vector; λ, the regularization parameter; T the maximum training epoch; $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ the training set.
output: Updated classifier of the learning model $h_T(\mathbf{x}) = \omega_T^\top \mathbf{x}$

for $t = 1$ **to** $T - 1$ **do**
  $t \leftarrow t + 1$;
  $\eta_t = \frac{1}{\lambda t}$; // `learning rate`
  Choose $i_t \in \{1, \ldots, n\}$ uniformly at random;
  if $y_t \langle \omega_t, \mathbf{x}_t \rangle < 1$ **then**
    $\omega_t \leftarrow (1 - \eta_t \lambda)\omega_t + \eta_t y_t \mathbf{x}_t$;
  else
    $\omega_t \leftarrow (1 - \eta_t \lambda)\omega_t$;
  end if
  $\omega_{t+1} \leftarrow \min\left(1, \frac{1}{\sqrt{\lambda}\|\omega_{t+1}\|}\right)\omega_{t+1}$; //`projection on` L_2 `norm`
end for
return: ω_T;

Evaluation of the upper bounds. Now to compare the upper bounds obtained by Bercu and Touati (2019) on the average risk of with the ones given in Cesa-Bianchi and Gentile (2008), we proceed as follows: we set $\delta = 10^{-5}$. For each regularization parameter $\lambda \in \{0.01, 0.1, 1, 10, 100\}$ we run a PEGASOS algorithm for different maximum numbers of epochs $T \in \{5, 50, 250, 500, 1000\}$. In Figure 2.18, we plot the upper bounds on the average risk as a function of the epochs. As it can be seen, the ratio between the two bounds is approximately 2 for a large T while it greater than 20 for a small epochs. A further interesting observation in this experiment consists in the fact that Cesa-Bianchi's upper bounds needs a large number of epochs to reach small values for instance 2, whereas the upper bound in this work can get this value for only small epochs.

The significant gap between our bounds and that of, motivates the development of new risk tail bounds for online learning algorithms. The first stone in this edifice will be the **Corollary 2.14**.

2.4 Two keystone lemmas

Our first lemma deals with a sharp upper bound on the Hermite generating function associated with a centered random variable X.

Lemma 2.17. *Let X be a square integrable random variable with zero mean and*

variance σ^2. *For all* $t \in \mathbb{R}$, *denote*

$$L(t) = \mathbb{E}\Big[\exp\Big(tX - \frac{at^2}{2}X^2\Big)\Big] \tag{2.46}$$

with $a > 1/8$. *Then, for all* $t \in \mathbb{R}$,

$$L(t) \leqslant 1 + \frac{b(a)t^2}{2}\sigma^2 \qquad \textit{where} \qquad b(a) = \frac{2a(1 - 2a + 2\sqrt{a(a+1)})}{8a - 1}. \tag{2.47}$$

Proof. In order to simplify the notation, denote $b = b(a)$. The proof of Lemma 2.17 relies on the following Hermite inequality, see also Proposition 12 in Delyon (2009) for the special value $a = 1/3$. For all $x \in \mathbb{R}$, we have

$$\exp\Big(x - \frac{ax^2}{2}\Big) \leqslant 1 + x + \frac{bx^2}{2}. \tag{2.48}$$

As a matter of fact, let

$$\varphi_a(x) = \log\Big(1 + x + \frac{bx^2}{2}\Big) - x + \frac{ax^2}{2}. \tag{2.49}$$

It is of course necessary to assume that $b > 1/2$ which ensures that $1 + x + bx^2/2$ is positive whatever the value of x is. We clearly have

$$\varphi_a'(x) = \Big(1 + x + \frac{bx^2}{2}\Big)^{-1} x P_{a,b}(x), \tag{2.50}$$

where the second degree polynomial $P_{a,b}$ is given by

$$P_{a,b}(x) = \frac{abx^2}{2} + \frac{(2a - b)x}{2} + a + b - 1.$$

Hereafter, assume that $a > 1/8$ and $b \neq 1 - a$. The unique positive root of the discriminant of $P_{a,b}$ is gven by $b = b(a)$. Consequently, as $\varphi_a'(0) = 0$ and $\varphi_a(0) = 0$, we deduce from (2.50) that the function φ_a reaches its minimum at $x = 0$ and we find that for all $x \in \mathbb{R}$, $\varphi_a(x) \geqslant 0$ which immediately leads to (2.48). Therefore, we obtain from (2.48) that for all $t \in \mathbb{R}$,

$$L(t) = \mathbb{E}\Big[\exp\Big(tX - \frac{at^2}{2}X^2\Big)\Big] \leqslant \mathbb{E}\Big[1 + tX + \frac{bt^2X^2}{2}\Big] = 1 + \frac{bt^2}{2}\sigma^2,$$

which is exactly what we wanted to prove. □

Our second exponential supermartingale lemma is as follows.

Lemma 2.18. *Let* (M_n) *be a locally square integrable real martingale. For all* $t \in \mathbb{R}$ *and* $n \geqslant 0$, *denote*

$$V_n(t) = \exp\Big(tM_n - \frac{at^2}{2}[M]_n - \frac{b(a)t^2}{2} <M>_n\Big) \tag{2.51}$$

with $a > 1/8$. *Then,* $(V_n(t))$ *is a positive supermartingale such that* $\mathbb{E}[V_n(t)] \leqslant 1$.

Proof. The proof follows from Lemma 2.17 together with standard arguments, see Bercu and Touati (2008) page 1860. □

2.5 Proofs of the main results

Proof of Theorem 2.1. For any positive x and y, let $A_n = \Big\{|M_n| \geqslant x, aS_n(a) \leqslant y\Big\}$. We have the decomposition $A_n = A_n^+ \cup A_n^-$ where $A_n^+ = \Big\{M_n \geqslant x, aS_n(a) \leqslant y\Big\}$ and $A_n^- = \Big\{M_n \leqslant -x, aS_n(a) \leqslant y\Big\}$. It follows from Markov's inequality together with Lemma 2.18 that for all positive t,

$$\begin{aligned}\mathbb{P}(A_n^+) &\leqslant \mathbb{E}\Big[\exp\Big(tM_n - tx\Big)\mathrm{I}_{A_n^+}\Big] \leqslant \mathbb{E}\Big[\exp\Big(tM_n - \frac{t^2}{2}aS_n(a)\Big)\exp\Big(\frac{t^2}{2}aS_n(a) - tx\Big)\mathrm{I}_{A_n^+}\Big],\\ &\leqslant \exp\Big(\frac{t^2 y}{2} - tx\Big)\mathbb{E}[V_n(t)] \leqslant \exp\Big(\frac{t^2 y}{2} - tx\Big).\end{aligned}$$

Hence, by taking the optimal value $t = x/y$ in the above inequality, we find that

$$\mathbb{P}(A_n^+) \leqslant \exp\Big(-\frac{x^2}{2y}\Big).$$

We also obtain the same upper bound for $\mathbb{P}(A_n^-)$ which ensures that

$$\mathbb{P}(A_n) \leqslant 2\exp\Big(-\frac{x^2}{2y}\Big). \tag{2.52}$$

Finally, inequality (2.52) clearly leads to (2.7) replacing y by ay. □

Proof of Theorem 2.4. For any positive x and y:
let $B_n = \Big\{|M_n| \geqslant xS_n(a), S_n(a) \geqslant y\Big\} = B_n^+ \cup B_n^-$
where $B_n^+ = \Big\{M_n \geqslant xS_n(a), S_n(a) \geqslant y\Big\}$ and $B_n^- = \Big\{M_n \leqslant -xS_n(a), S_n(a) \geqslant y\Big\}$.
Proceeding as in the proof of Theorem 2.1, we have that for all positive t such that $t < 2x/a$,

$$\mathbb{P}(B_n^+) \leqslant \mathbb{E}\Big[\exp\Big(tM_n - txS_n(a)\Big)\mathrm{I}_{B_n^+}\Big] \leqslant \mathbb{E}\Big[\exp\Big(tM_n - \frac{t^2}{2}aS_n(a)\Big)\exp\Big(\frac{t}{2}(ta-2x)S_n(a)\Big)\mathrm{I}_{B_n^+}\Big]$$

Thus:

$$\mathbb{P}(B_n^+) \leqslant \exp\Big(\frac{t}{2}(ta-2x)y\Big)\mathbb{E}[V_n(t)] \leqslant \exp\Big(\frac{t}{2}(ta-2x)y\Big). \tag{2.53}$$

Consequently, we find from (2.53) with the particular choice $t = x/a$ that

$$\mathbb{P}(B_n^+) \leqslant \exp\Big(-\frac{x^2 y}{2a}\Big). \tag{2.54}$$

The same upper bound holds for $\mathbb{P}(B_n^-)$ which clearly implies (2.8). Furthermore, for any positive x, let $C_n = \Big\{|M_n| \geqslant xS_n(a)\Big\} = C_n^+ \cup C_n^-$ where $C_n^+ = \Big\{M_n \geqslant$

$xS_n(a)\Big\}$ and $C_n^- = \Big\{M_n \leqslant -xS_n(a)\Big\}$. By Holder's inequality, we have for all positive t and $q > 1$,

$$\begin{aligned}
\mathbb{P}(C_n^+) &\leqslant \mathbb{E}\Big[\exp\Big(\frac{t}{q}M_n - \frac{tx}{q}S_n(a)\Big)\mathrm{I}_{C_n^+}\Big], \\
&\leqslant \mathbb{E}\Big[\exp\Big(\frac{t}{q}M_n - \frac{t^2a}{2q}S_n(a)\Big)\exp\Big(\frac{t}{2q}(ta-2x)S_n(a)\Big)\mathrm{I}_{C_n^+}\Big], \\
&\leqslant \mathbb{E}\Big[\big(V_n(t)\big)^{1/q}\exp\Big(\frac{t}{2q}(ta-2x)S_n(a)\Big)\Big], \\
&\leqslant \Big(\mathbb{E}\Big[\exp\Big(\frac{tp}{2q}(ta-2x)S_n(a)\Big)\Big]\Big)^{1/p}. \qquad (2.55)
\end{aligned}$$

Consequently, as $p/q = p-1$, we can deduce from (2.55) with the optimal value $t = x/a$ that

$$\mathbb{P}(C_n^+) \leqslant \inf_{p>1}\Big(\mathbb{E}\Big[\exp\Big(-\frac{(p-1)x^2S_n(a)}{2a}\Big)\Big]\Big)^{1/p}.$$

We find the same upper bound for $\mathbb{P}(C_n^-)$, completing the proof of Theorem 2.4. □

Proof of Theorem 2.6. We already saw from 2.18 that for all $t \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\Big(tA_n - \frac{t^2}{2}B_n^2\Big)\Big] \leqslant 1$$

where $A_n = M_n$ and $B_n^2 = a[M]_n + b(a) < M >_n$. It means that the pair of random variables (A_n, B_n) safisties the canonical assumption in De la Peña and Pang (2009). Theorem 2.6 follows from Theorem 2.1 in Delyon (2009). □

Chapter 3

Online Learning:New Frontiers in risk tail bounds.

abstract

We prove, for an arbitrary online learning algorithm, new tight bounds for specific hypothesis selected from the ensemble generated by this algorithm. The main idea is essentially based on a suitable use of new concentration inequalities for martingales. This type of inequality has not yet been exploited in the field of online statistical learning more precisely to obtain risk bounds. The use of new concentration inequalities (Bercu-Touati2019, Bercu et al. 2015) allowed us to obtain drastically tighter bounds than those obtained previously by Cesa-Bianchi and Gentile. This theoretical realization, shows that the inequalities of Bercu-Touati potentially constitute a powerful tool for the improvement of more sophisticated bounds in this field. To demonstrate the relevance of our proposed risk tail bounds, we conduct several numerical experiments on both synthetic and real data.

3.1 Introduction

This chapter is intrinsically linked to Bercu and Touati (2019), insofar we use theese results to propose a clear improvement of the risk tail bounds proposed by Cesa-Bianchi and Gentile (2008).

It is important to specify that these risk tail bounds are valid for any algorithm. In this direction, no improvement has been made so far since 2008. The main idea allowing the improvement of the previously mentioned results rests on the use of new concentration inequalities for martingales. Until then in the literature, this type of result was built on the basis of old classical inequalities Azuma (1967), Freedman (1975).

We will see that through the work of Bercu and Touati (2019), we considerably

simplify the analysis made by Cesa-Bianchi and Gentile (2008) and we obtain drastically tighter risk tail bounds that we have evaluated with several recent algorithms. We analyze the risk of models selected from the ensemble produced by training a learning algorithm incrementally on a sequence of independent and identically distributed (i.i.d.) data.

We adopt the same paradigm established by Cesa-Bianchi et al. (2004), to the extent that we analyze the underlying empirical process (associated to the online learner) through concentration inequalities for martingales. Instead of using the classic concentration inequalities, it is through the inequalities of Bercu and Touati (2019) that we control the deviation of the online-process. These inequalities through their parametrization, gives flexibility and guarantee the possibility of obtaining an optimal bound on the average risk linked to the algorithm used. Through an adaptation of **Theorem 1.6** and **Theorem 1.8**, we considerably simplify the previous analysis in Cesa-Bianchi and Gentile (2008) (Section III), we guarantee also the simplest possible penalisation. The chapter is organized as follows: We first introduce the problem to be considered, namely the binary classification in an online way, we then introduce the basic notations. Then we state the main results consisting in providing risk tail bounds for any online learning algorithm. We demonstrate the efficiency of our terminals via numerical simulations and we compare them to the reference work of Cesa-Bianchi and Gentile (2008) Finally a section is dedicated to the detailed proof of mathematical results.

3.2 Problem setup

Binary classification. Given a sequence of training examples $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_t \in \mathcal{X} = \mathbb{R}^d$ is a d-dimensional instance representing the features and $y_t \in \mathcal{Y} = \{-1, +1\}$ for binary classification, $y_t \in \{0, 1\}^L$ for multi-class classification with L classes, and $y_t \in \mathbb{R}$ for regressions tasks, is the target label assigned to $\mathbf{x}_t$. Online learning algorithms operates on a sequence of data examples with time stamps. At each step t, the learner receives an incoming example $\mathbf{x}_t \in \mathcal{X}$ in a d-dimensional vector space, that is, $\mathcal{X} = \mathbb{R}^d$. It first attempts to predict the class label of the incoming instance, $\hat{y}_t = \mathtt{sgn}\{< \mathbf{w}_t \cdot \mathbf{x}_t >\} \in \mathcal{Y}$, and $\mathcal{Y} = \{-1, +1\}$ for binary classification tasks. ($\mathbf{w}_t = h_{t-1}(\mathbf{x}_t)$). After making the prediction, the true label $y_t \in \mathcal{Y}$ is revealed, and the learners then computes the loss $\ell(y_t, \hat{y}_t)$ based on some criterion to measure the difference between the learner's prediction and the revealed true label y_t.

For any classifier h we define the risk (resp.) the empirical risk associated to the training sample $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ as follows:

$$R(h) = \mathbb{E}[\ell(h(X), Y)] \text{ resp. } \widehat{R}(h) = \frac{1}{n}\sum_{t=1}^{n} \ell(h(X_t), Y_t).$$

In online learning setting, for a sequence of learners $h_0, h_1, \ldots, h_{n-1}$, we define the

following empirical risk $\widehat{R}_n = \frac{1}{n}\sum_{t=1}^{n} \ell(h_{t-1}(X_t), Y_t))$. For a fixed $t \in \{0, \ldots, n-1\}$ we define also $\widehat{R}_n^{(t)} = \frac{1}{n-t}\sum_{i=t+1}^{n} \ell(h_t(X_i), Y_i))$ which corresponds to the empirical risk of the classifier h_t on the reamining samples $(X_{t+1}, Y_{t+1}), \ldots, (X_n, Y_n)\}$. For a fixed $\delta \in]0, 1]$, we define the penalized empirical risk denoted by

$$\mathrm{PER}_{n,\delta}(h_t) = \widehat{R}_n^{(t)} + \mathrm{pen}_\delta(\widehat{R}_n^{(t)}, t)$$

and

$$\widehat{h} = \mathrm{argmin}_{0 \leqslant t < n}\{\mathrm{PER}_{n,\delta}(h_t)\}$$

We introduce also the following functions wich are useful to our first result:

$\forall (r, t) \in]0, 1[\times\mathbb{R}_+$, and $B = \log(\frac{n.(n+2)}{\delta})$

$$\mathrm{pen}_\delta(r, t) = \sqrt{\frac{(1-r^2)}{\log(\frac{1}{r})}.\frac{B}{n-t}}$$

$$\Psi_B(r, t) = \sqrt{\frac{13}{5}\frac{B.r}{n-t}}$$

$$F_\delta(r, t) = 2.\,\mathrm{pen}_\delta(r + \Psi_B(r, t), t)$$

3.3 Main results

Lemma 3.1. *Let $h_0, \ldots, h_{n-1}$, be the ensemble of hypotheses generated by an arbitrary online algorithm working with a loss ℓ having values in $[0, 1]$. Then, the determenistic hypothesis $\widehat{h}$ satifies*

$$R(\widehat{h}) \leqslant \min_{0 \leqslant t < n}\{R(h_t) + F_\delta(R(h_t), t)\},$$

with probability at least $1 - \delta$.

Remark 3.2. *We simplify, as much as possible, the analysis made by Cesa-Bianchi and Gentile (2008) (in the section III).*
While keeping the same, some technical tools , we use more sophisticated concentration inequalities and we obtain a much more smaller penalization and a shorter and more elegant demonstration.
In this section, we show how to choose a hypothesis, in a deterministic way in order to obtain a tight risk bound. Although based on a martingale underlying structure, the bound that we construct for this hypothesis is not directly comparable to the bound for the average random hypothesis.

$n \geqslant 1$ and $\delta \in]0, 1]$ being fixed beforehand,let us introduce the following function:

$\forall t \in \{0, ..., n-1\}$ and $\forall x \geqslant 0$ $(C = \log(\frac{2n.(n+2)}{\delta})$.

$$g_t(x) = x + \sqrt{\frac{C}{3}}(\sqrt{1 + 4.(x + \Psi_C(x,t)) + (x + \Psi_C(x,t)^2)})\sqrt{\frac{\log(n-t)}{n-t}}$$

$\forall a \in]\frac{1}{8}, \frac{9}{16}]$ and $\forall x \in]0, 1]$:

$$c(a) = \frac{2(1 - 2a + 2\sqrt{a(a+1)})}{8a - 1}$$

$$\Phi_a^{-1}(x) = x + \frac{ac(a)\log(2n/\delta)}{n} + \sqrt{\frac{a\log(2n/\delta)}{n}\left(2 + 2c(a)x + \frac{ac^2(a)\log(2n/\delta)}{n}\right)}$$

We are now inclined to introduce our main result:

Theorem 3.3. *Let $h_0, \ldots, h_{n-1}$, be the ensemble of hypotheses generated by an arbitrary online algorithm working with a loss ℓ having values in $[0, 1]$. Then, as soon as $g_t\left(\Phi_a^{-1}(.)\right) \leqslant 1$ $(^*)$ The determenistic hypothesis $\widehat{h}$ satifies*

$$R(\widehat{h}) \leqslant \min_{0 \leqslant t < n} g_t\left(\Phi_a^{-1}(\widehat{R}_n^{(t)})\right)$$

with probability at least $1 - \delta$.

We would like to point out that the quantity $c(a)$ and have already been defined in the previous chapter 2.6. Φ_a^{-1} looks a lot like the one used in Chapter 2 2.44, we just replace δ by $\frac{\delta}{2n}$.

Each g_t is monotonically increasing on $[0, 1]$. The best bound is obtained when $t = 0$. The condition $(^*)$ guarantees the effectiveness of the bounds provided i.e (that they are smaller than 1).

At a fixed $(1 - \delta).100\%$ confidence level, $(^*)$ depends exclusively on the size of the sample and the accurancy of the used agorithm.

Obviously, we will study only algorithms performing better than a naive classifier, i.e having an accurancy greater than 50% (empirical risk smaller than 0.5).

Remark 3.4. *For a fixed 99% confidence threshold i.e $\delta = 0.01$, the condition $(^*)$ is satisfied for any non-naive algorithm as long as the sample size is greater than 500. Technically, this amounts to numerically solving the following equation*

$$x = \Phi_a(g_0^{-1}(1))$$

(x represents the empirical risk. The most pessimistic case being that for an a in a neighborhood of $\frac{1}{8}$ for example $a = 0.13$). If we set the level of confidence and the risk not to be exceeded, that is to say:

$$0.5 > \Phi_a(g_0^{-1}(1))$$

and $\delta = 0.01$ *The only unknown quantity is therefore the sample size, and as long as the sample size is greater than 500 we guarantee the efficiency of our bounds for any non-naive online algorithm.*

Remark 3.5. *Inequality 2.37 can be improved by using Theorem 1.6 considering*

$$M_n = n.\Big(\widehat{R}(h) - \widehat{R}_n)\Big), \tag{3.1}$$

no longer as a martingale but as a sum of random variables bounded by 1. We thus obtain a smaller bound than $\Phi_a^{-1}(\widehat{R}_n)$ *using numerical inversion.* ***Each*** g_t ***being monotonically increasing on*** $[0, 1]$***.*** *We therefore digitally obtain an improved version of the bounds. We will call this bound,* ***improved optimal bound****. (It will be gray in the graphics).*
In all cases, this bound surpasses all the others and it is efficient from a small sample size which justifies its importance although it is built with a numerical inversion.

We highlight it only in the last example so that it is not redundant.

3.4 Numerical experiments

We conduct numerical experiments to demonstrate the relevance of our proposed risk tail bounds. We evaluate these bounds for the following algorithms: Passive Aggressive, Exact Soft Confidence-Weighted Learning, Multiple Layer Perceptron. We first evaluate our bounds on simulated synthetic data, We deal with two classic cases in binary classification, linearly and non-linearly separable data respectively. What is important is to be able to evaluate our bounds for small, intermediate and high values of the empirical risk.
The choice of algorithms and simulations is based on this important criterion in order to evaluate the bounds in the following situations: bad classification, intermediate classification, almost perfect classification.
Finally, we evaluate our bounds on small data sets to challenge their effectiveness even when the sample size is low. Indeed a bound greater than 1, does not provide any relevant information (Beforehand, we can say deterministically that the empirical risk is always smaller than 1.)
The results of Bercu and Touati (2019) allow us to take advantage of a great flexibility and to provide an optimal bound for each algorithm which is in itself a major advance in this field.
We first present the data used, then we will give a short presentation of the algorithms used with a commentary on the plots containing the Touati bounds and those of Cesa-Bianchi and Gentile (2008).

3.4.1 Experimental protocol

We fix for the simulations a size of sample equal to 10000. This is more than sufficient in the framework of non asymptotic bounds. With this sample size

we can demonstrate the speed of convergence of our bounds and their precision. We compare our bounds to those of Cesa-Bianchi and Gentile (2008) by running each online learning algorithm on linearly and non linearly separable data. We have chosen two linear classifiers namely Online Passive-Aggressive Algorithm and Exact Soft Confidence-Weighted Learning and a non-linear classifier: Online Perceptron. This choice will be profitable for us to show the effectiveness of our bounds compared to those established in the literature for high and low values of the empirical risk. Our goal is not to compare the accurancy of algorithms but to highlight the drastic improvement of the bounds of Cesa-Bianchi and Gentile (2008) in several situation. We finally evaluate our bounds by running the algortihms on real data set with small size this allows us to see that the bounds Cesa-Bianchi and Gentile (2008) are not at all efficient (i.e. always much larger than 1) for small databases unlike ours which are effective in this case. Finally thanks to an improvement of inequality 2.37, we assess an improved optimal bound which substantially improves all the results. The only drawback of this new bound is that it is obtained via a numerical inversion and does not have an analytical expression. **Linearly separable data** We set $n = 10000$ and for each $y_i \in \{-1, +1\}$, we draw samples $x_i \sim \mathcal{N}((0,0)^\top, \left(\begin{smallmatrix} 1.1 & 0.1 \\ 0.1 & 1.1 \end{smallmatrix}\right))$ for $i \in \{1, \ldots, 5000\}$ and $x_i \sim \mathcal{N}((3,3)^\top, \left(\begin{smallmatrix} 1.2 & 0.2 \\ 0.2 & 1.2 \end{smallmatrix}\right))$ for $i \in \{5000, 10000\}$.

We notice via the following plot that the data are not perfectly linearly separable, this avoids us a perfect classification with a zero empirical risk.

We notice via the following plot that the data are not perfectly linearly separable, this avoids us a perfect classification with a zero empirical risk.

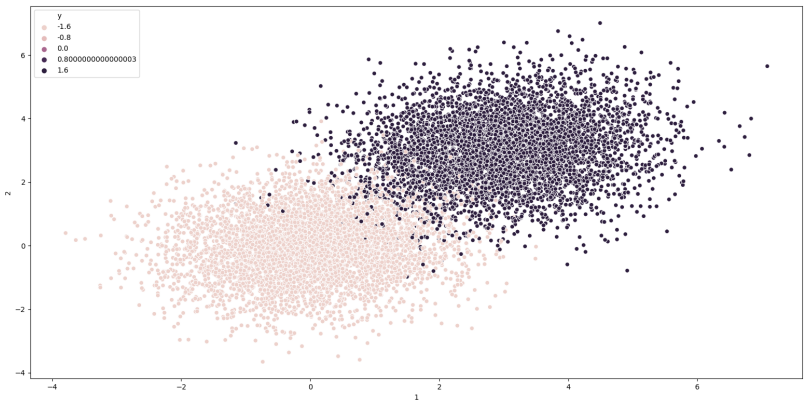


Figure 3.1 – Linearly separable simulations

Non linearly separable data

Algorithm 2 Simulation of non linearly separable data

input: U and V Two random variables distributed according to the standard uniform distribution.
output: data frame containing features X,Y and a binary variable of interest Z.
$T = 10^4$
$R \leftarrow 40 * U$;
$F \leftarrow which\ R > 20$; // `Far points`
$R[F] \leftarrow R[F] * 1.2$;
$R[\bar{F}] \leftarrow R[\bar{F}] * 1.1$; // $\bar{F}$ `is the complement (set theory) of` F.
$\Theta \leftarrow 2\pi * V$;
$X \leftarrow R\cos\theta$;
$Y \leftarrow R\sin\theta$;
if $X^2 + Y^2 \leqslant 400$ **then**
  $Z \leftarrow 1$
else
  $Z \leftarrow -1$
end if

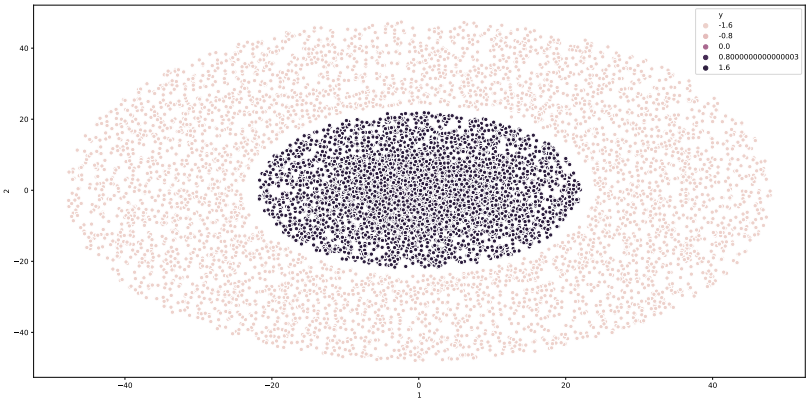


Figure 3.2 – Non linearly separable simulations

Ionosphere data set We downloaded this database on, UCI Machine Learning Repository. It contains 34 attributes and a binary response variable. The size of this database is relatively small which is particularly challenging to assess the effectiveness of bounds.
The target variable is intrinsically linked to a physical phenomenon of radar measurement of the ionosphere. Radar returns from the ionosphere are classified as either suitable for further analysis or not.

In the literature, a reference article Sigillito et al. (1989) confirms the accurancy of neural networks, which we will highlight in our simulations, but the main interest remains the efficiency of bounds evaluated on small data sets.

Figure 3.3 – Ionosphere data

Breast Cancer data set It's a very popular database on scikit-learn Python, having 30 features and a binary target variable which classifies tumors into malignant and benign. As previously described the main interest remains more the evaluation of bounds than the accurancy.

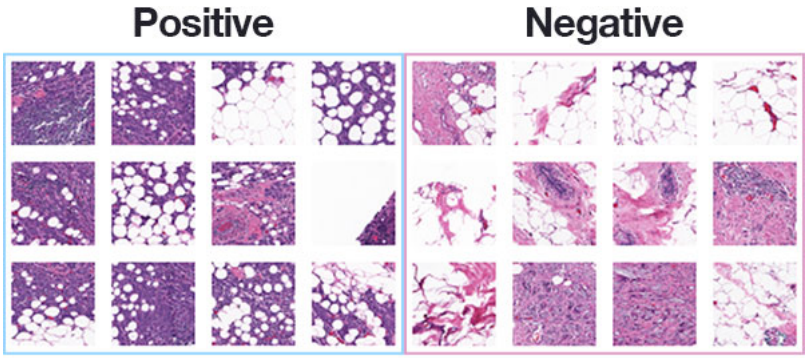


Figure 3.4 – Breast cancer classification

3.4.2 Passive Aggressive algorithm

The Passive Aggressive algorithm, developped by Crammer et al. (2006) is a first order first order online learning algorithm designed for both regression and classification problems. The update of hypothesis is based on an optimization problem under constraints. We base ourselves on the same notations introduced

previously, to know:

$$\forall t \in \{1, ..., n\}\ \mathbf{z}_t = (\mathbf{x}_t, y_t)$$

is an example of an instance and his target value. Before describing the algorithm, we recall the three basic notions on which it is based:

— The notion of discrepancy measured for a hypothesis w as $\delta(\mathbf{w}; \mathbf{z}_t) = -y_t < \mathbf{w}.\mathbf{x}_t >$.

— The notion of Hinge Loss to measure the gap between the prediction and the target value.

$$\ell_\varepsilon(\mathbf{w}; \mathbf{z}_t) = [\delta(\mathbf{w}; \mathbf{z}_t) - \varepsilon]_+ = \max\{0, \delta(\mathbf{w}; \mathbf{z}_t) - \varepsilon\}$$

(ε is an accuracy parameter).

— The notion of realizability: We assume the existence of hypothesis $\mathbf{w}^*$ achieving zero loss over the sequence.

The operation of the algorithm is as follows:

— Each example defines a set of consistent hypothesis $C_\varepsilon(\mathbf{z}_t) = \{\mathbf{w} \mid \delta(\mathbf{w}; \mathbf{z}_t) \leqslant \varepsilon\}$

— The updated hypothesis $\mathbf{w}_{t+1}$ is set to be the projection of $\mathbf{w}_t$ onto $C_\varepsilon(\mathbf{z}_t)$. i.e $\mathbf{w}_{t+1} = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w} - \mathbf{w}_t\|$ s.t. $\mathbf{w} \in C_\varepsilon(\mathbf{z}_t)$

Algorithm 3 Passive Aggressive algorithm (PA)

INPUT:: Insensitivity parameter $\varepsilon > 0$
INITIALIZE $\mathbf{w}_0 = (0, ..., 0)$
for $t \in \{1, ...n\}$ **do**
  Get a new example $\mathbf{z}_t$
  Suffer loss $\ell_\varepsilon(\mathbf{w}_t; \mathbf{z}_t)$
  if $\ell_\varepsilon(\mathbf{w}_t; \mathbf{z}_t) > 0$ **then**
    1/Compute direction $\mathbf{v}_t = y_t.\mathbf{x}_t$
    2/Compute $\tau_t = \frac{\ell_\varepsilon(\mathbf{w};\mathbf{z}_t)}{||\mathbf{v}_t||^2}$
    3/ Update $\mathbf{w}_t + 1 = \mathbf{w}_t + \tau_t.\mathbf{v}_t$
  end if
end for

To ensure that PA is able to handle non-separable instances and to guarantee robustness, two variants are proposed. The modification is based on the introduction of two penalties, more precisely a linear and a quadratic penalty, leading to the following two formulations of soft-margin PA algorithms.

$$\mathbf{w}_{t+1}^{\text{PA}-1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 + C\ell_\varepsilon(\mathbf{w}; \mathbf{z}_t)$$

$$\mathbf{w}_{t+1}^{\text{PA}-\text{II}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 + C\ell_\varepsilon(\mathbf{w}; \mathbf{z}_t)^2$$

Where $C > 0$ is a parameter to control aggressiveness of PA. The resulting direction updates to the soft-margin PA algorithms have the same form as that of the original algorithm but they have different directions:

$$\mathbf{v}_t^{\text{PA-I}} = \min\left\{C, \frac{\ell_\varepsilon(\mathbf{w}_t; \mathbf{z}_t)}{\|\mathbf{x}_t\|^2}\right\}, \mathbf{v}_t^{\text{PA-II}} = \frac{\ell_\varepsilon(\mathbf{w}_t; \mathbf{z}_t)}{\|\mathbf{x}_t\|^2 + \frac{1}{2C}}$$

.

We run this algorithm on the simulations and the data sets presented previously. We already know that its performance will be satisfactory for linearly separable simulation; the inseparable case is very difficult to manage for the PA algorithm insofar as the border separating the target values is a circle. PA algorithm will behave like a naive classifier in this case. We will have treated then the intermediate and high case for empirical risk. We plot on the same graph, our optimal bound, our least efficient bound, that of Cesa-Bianchi and Gentile (2008) and of course the evolution of the empirical risk.

We notice on the one hand that for all the situations, our bounds are much more precise than those of Cesa-Bianchi and Gentile (2008) for any a even in the neighborhood of $\frac{1}{8}$, on the other hand, our bounds converge faster towards the minimum value of the empirical risk. We will comment in more detail each plot, to highlight the improvoments.

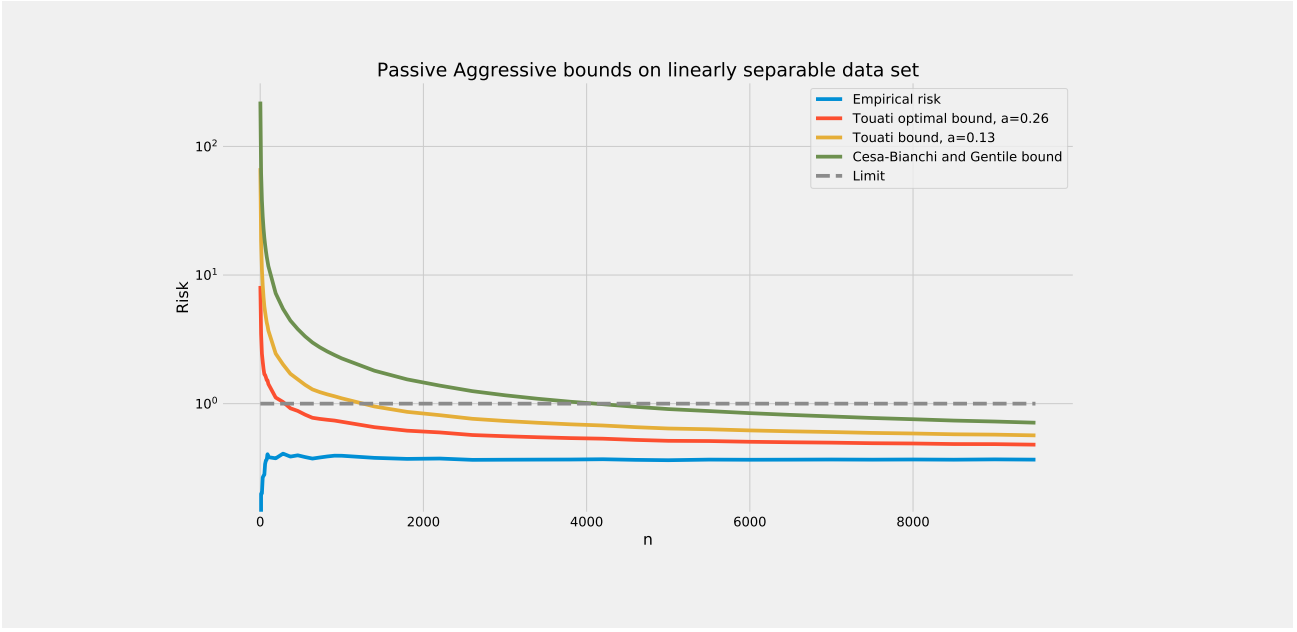


Figure 3.5 – Passive Aggressive bounds on linearly separable simulation

A more detailed analysis of this plot, provided that our bounds are efficient, i.e. smaller than 1 from **300** observations while those of Cesa-Bianchi and Gentile

(2008) are effective from **4900** observations. This is extremely important in terms of concrete applications.

Note also that for a very large sample size, the bound of Cesa-Bianchi and Gentile (2008) is not informative since it takes the value of **0.77** in the neighborhood of 10^4. This information is known in a deterministic manner, since the PA algorithm surpasses a naive classifier. Our bound is however effective and informs us, that with probability 99%, the empirical risk is less than **0.48**. The minimum value of the empirical risk reached by the algorithm is **0.35**.

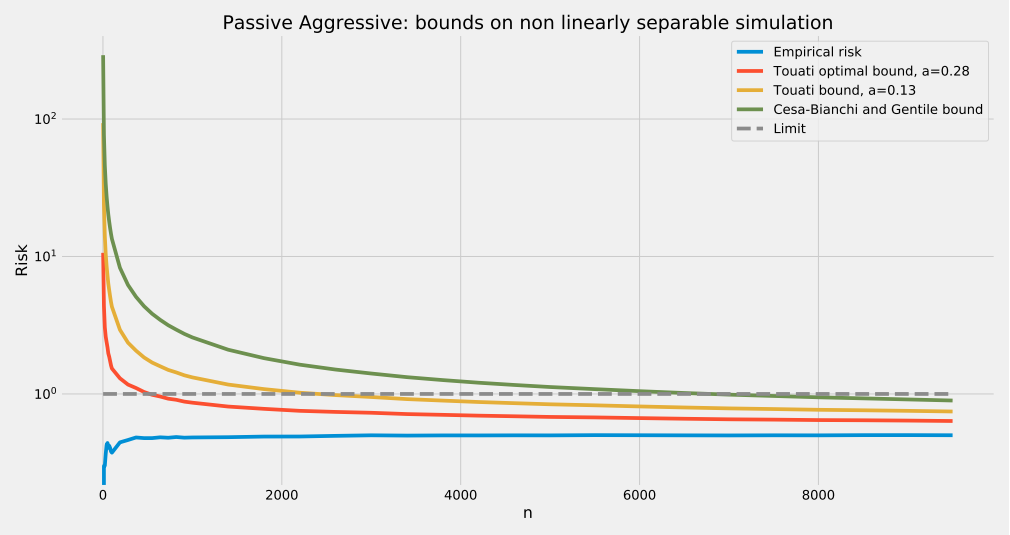


Figure 3.6 – Passive Aggressive bounds on non linearly separable simulation

We observe the same phenomenon for the nonlinearly separable case. However the performances are reduced because the emirical risk is very close to **0.5**, PA thus behaves practically like a naive classifier. Our bound is effective from **500** observations, that of Cesa-Bianchi and Gentile (2008) from **7000** observations. The minimum values of the bounds in the neighborhood of $n = 10^4$ are respectively **0.65** and **0.9**. Although being not informative in this case,the value provided by our bound remains relatively close to **0.5**. This case is obviously to be avoided in practice but from a theoretical point of view it is reasonably interesting to check out the quality of our bounds for a very bad classifier.

In the case of real data with a small number of observations, the bounds of Cesa-Bianchi and Gentile (2008) are inefficient for both a good or a bad classifier i.e the green curve is well above the dotted limit set at 1. Our bounds are not very informative when the classifier is bad and with a very small sample size.

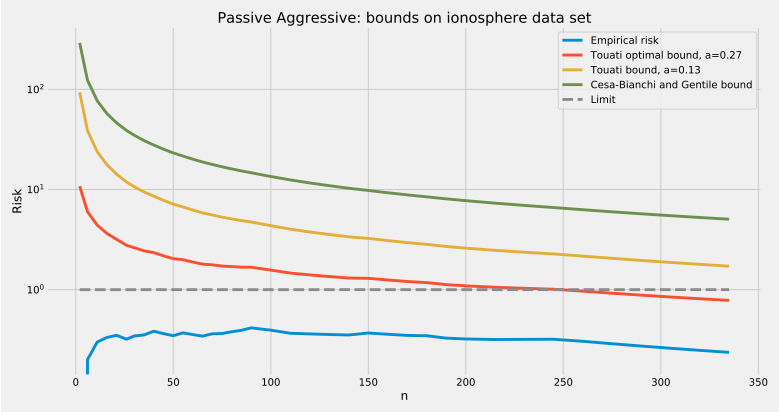


Figure 3.7 – Passive Aggressive bounds on ionosphere data set

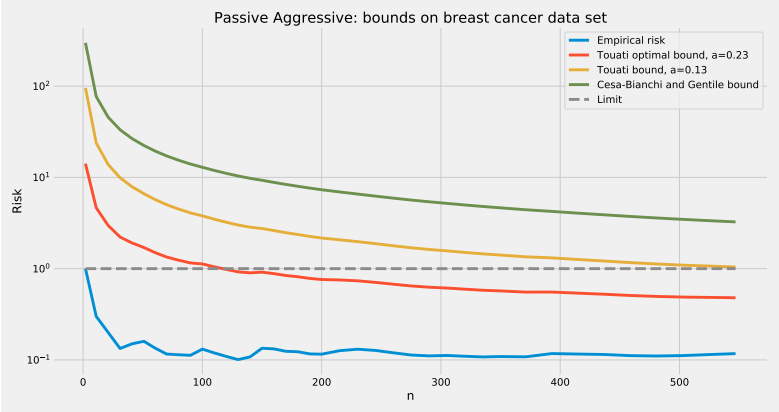


Figure 3.8 – Passive Aggressive bounds on breast cancer data set

3.4.3 Exact Soft Confidence-Weighted Learning (SCW)

This algorithm was developed by Wang et al. (2012) in order to better explore the underlying structure between features by assuming that the weights distribution is Gaussian. This improvement follows on from previous work, insofar as SCW has four fundamental properties that other confidence weighted algorithms Dredze et al. (2008) and Crammer et al. (2009) do not have:

— Large margin training
— Confidence weighting
— Aptitude to operate on non-separable data.
— Adaptive margin.

We present first, Confidence Weighted algorithms (CW) and the Adaptive Regularization of Weights then we explain in detail the functioning of the SCW agorithm.

The theoretical results concerning the performance of this method are fully explained in Wang et al. (2012).
This algorithm outperforms PA algorithm in terms of accurancy and has more guarantees on computational efficiency.

Confidence-Weighted Learning (CW) CW algorithms were developed by Dredze et al. (2008), the main assumption is that the weights i.e. learners are distributed according to a normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The updation of the weight's distribution takes place by minimizing the Kullback-Leibler divergence between the new weight distribution and the one before her, while ensuring that the probability of misclassification is below a threshold set beforehand.
This algorithm is therefore formalized as an optimization problem under constraints as follows:

$$\left(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1}\right) = \arg\min_{\boldsymbol{\mu},\Sigma} D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}\left(\boldsymbol{\mu}_t, \Sigma_t\right)) \quad \text{s.t. } \mathbb{P}\left[y_t \left\langle \mathbf{w} \cdot \mathbf{x}_t \right\rangle \leqslant 0\right] \leqslant \delta$$

where $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

The solution associated with this problem is of the closed form, it is formulated as follows:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \alpha_t y_t \Sigma_t \mathbf{x}_t \quad \Sigma_{t+1} = \Sigma_t - \beta_t \Sigma_t \mathbf{x}_t^T \mathbf{x}_t \Sigma_t \tag{**}$$

The updating coefficients are calculated as follows:

$$\alpha_t = \max\left\{0, \frac{1}{b_t\gamma}\left(-m_t\psi + \sqrt{m_t^2\frac{\phi^4}{4} + b_t\phi^2\gamma}\right)\right\}$$
$$\beta_t = \frac{\alpha_t\phi}{\sqrt{a_t} + b_t\alpha_t\phi}$$

where $a_t = \frac{1}{4}\left(-\alpha_t b_t \phi + \sqrt{\alpha_t^2 b_t^2 \phi^2 + 4b_t}\right)^2, b_t = \mathbf{x}_t^T \Sigma_t \mathbf{x}_t, m_t = y_t \left\langle \boldsymbol{\mu}_t \cdot \mathbf{x}_t \right\rangle, \phi = \Phi^{-1}(\delta)$
(Φ is the cummulative function of the normal distribution) $, \psi = 1 + \frac{\phi^2}{2}$, and $\gamma = 1 + \phi^2$.

Adaptive Regularization of Weights

$$\left(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1}\right) = \arg\min_{\boldsymbol{\mu},\Sigma} D_{KL}\left(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}\left(\boldsymbol{\mu}_t, \Sigma_t\right)\right) + \frac{1}{2C}\ell^2\left(\boldsymbol{\mu}; (\mathbf{x}_t, y_t)\right) + \frac{1}{2C}\mathbf{x}_t^T \Sigma_t \mathbf{x}_t$$

where $\ell^2(\boldsymbol{\mu}; (\mathbf{x}_t, y_t)) = (\max\{0, 1 - y_t \langle \boldsymbol{\mu} \cdot \mathbf{x}_t \rangle\})^2$ and C is a regularization parameter. The closed-form solution of the optimization problem is close to the previous one of $^{(**)}$ but with different updating coefficients:

$$\alpha_t = \ell(\boldsymbol{\mu}_t; (\mathbf{x}_t, y_t))\, \beta_t,\ \beta_t = \frac{1}{\mathbf{x}_t^T \Sigma_t \mathbf{x}_t + C}$$

Soft Confidence-Weighted Learning This alogorithm eloborated by Wang et al. (2012), was set up to overcome the shortcomings of the two approaches presented before. Following the same problem settings of the Confidence-Weighted learning, the weight vector w follows the Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ. Notice that the probability constraint in the CW learning, i.e., $\mathbb{P}[y_t \langle \mathbf{w} \cdot \mathbf{x}_t \rangle \leqslant 0] \leqslant \delta$ can be rewritten as

$$y_t (\boldsymbol{\mu} \cdot \mathbf{x}_t) \leqslant \phi \sqrt{\mathbf{x}_t^\top \Sigma \mathbf{x}_t}$$

where $\phi = \Phi^{-1}(\delta)$. Further, the loss function considered is as follows:

$$\ell^\phi(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\mathbf{x}_t, y_t)) = \max\left(0, \phi \sqrt{\mathbf{x}_t^\top \Sigma \mathbf{x}_t} - y_t \boldsymbol{\mu} \cdot \mathbf{x}_t\right)$$

It is easy to verify that satisfying the probability constraint (i.e., $y_t (\boldsymbol{\mu} \cdot \mathbf{x}_t) \leqslant \phi \sqrt{\mathbf{x}_t^\top \Sigma \mathbf{x}_t}$ for any $\phi > 0$) is equivalent to satisfying $\ell^\phi(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\mathbf{x}_t, y_t)) = 0$ Therefore, the optimization problem of the original CW can be re-written as follows

$$\begin{aligned}\left(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1}\right) = &\arg\min_{\boldsymbol{\mu}, \Sigma} D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \| \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)) \\ &\text{s.t. } \ell^\phi(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\mathbf{x}_t, y_t)) = 0, \phi > 0\end{aligned}$$

To overcome the above limitation of the CW learning problem, Wang et al. (2012) propose a Soft Confidence-Weighted (SCW) learning method, which aims to soften the aggressiveness of the weights updating strategy. The optimization of SCW for learning the soft-margin classifiers is formulated as follows:

$$\begin{aligned}\left(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1}\right) = &\arg\min_{\boldsymbol{\mu}, \Sigma} D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \| \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)) \\ &+ C\ell^\phi(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\mathbf{x}_t, y_t))\end{aligned} \quad \textbf{(P-SCW-I)}$$

where C is a parameter to tradeoff the passiveness and aggressiveness. The above formulation of the Soft Confidence-Weighted algorithm is called "SCW-I" for short. Similar to the variant of PA, the above formulation can be enhanced by employing a squared penalty, leading to the second formulation of SCW learning (denoted as "SCW-II" for short):

$$\begin{aligned}\left(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1}\right) = &\arg\min_{\boldsymbol{\mu}, \Sigma} D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \| \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)) \\ &+ C\ell^\phi(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\mathbf{x}_t, y_t))^2\end{aligned} \quad \textbf{(P-SCW-II)}$$

We recall the two results to build the algorithm:

Proposition 1: Wang et al. (2012)
The closed-form solution of the optimization problem **P-SCW-I** is expressed as follows:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \alpha_t y_t \Sigma_t \mathbf{x}_t, \Sigma_{t+1} = \Sigma_t - \beta_t \Sigma_t \mathbf{x}_t^T \mathbf{x}_t \Sigma_t$$

where the updating coefficients are as follows:

$$\alpha_t = \min\left\{C, \max\left\{0, \frac{1}{v_t\zeta}\left(-m_t\psi + \sqrt{m_t^2\frac{\phi^4}{4} + v_t\phi^2\zeta}\right)\right\}\right\}$$
$$\beta_t = \frac{\alpha_t\phi}{\sqrt{u_t} + v_t\alpha_t\phi}$$

where $u_t = \frac{1}{4}\left(-\alpha_t v_t\phi + \sqrt{\alpha_t^2 v_t^2\phi^2 + 4v_t}\right)^2, v_t = \mathbf{x}_t^T\Sigma_t\mathbf{x}_t, m_t = y_t\left(\boldsymbol{\mu}_t \cdot \mathbf{x}_t\right), \phi = \Phi^{-1}(\delta), \psi = 1 + \frac{\phi^2}{2}$ and $\zeta = 1 + \phi^2$

Proposition 2: Wang et al. (2012) The closed-form solution of the optimization problem **P-SCW-II** is:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \alpha_t y_t \Sigma_t \mathbf{x}_t, \Sigma_{t+1} = \Sigma_t - \beta_t \Sigma_t \mathbf{x}_t^T \mathbf{x}_t \Sigma_t$$

The updating coefficients are as follows:

$$\alpha_t = \max\left\{0, \frac{-\left(2m_t n_t + \phi^2 m_t v_t\right) + \gamma_t}{2\left(n_t^2 + n_t v_t \phi^2\right)}\right\}$$
$$\beta_t = \frac{\alpha_t\phi}{\sqrt{u_t} + v_t\alpha_t\phi}$$

where $\gamma_t = \phi\sqrt{\phi^2 m_t^2 v_t^2 + 4n_t v_t\left(n_t + v_t\phi^2\right)}$, and $n_t = v_t + \frac{1}{2C}$

Algorithm 4 SCW learning algorithms (SCW)

INPUT:: parameters $C > 0, \delta > 0$
INITIALIZE $\boldsymbol{\mu}_0 = (0, \ldots, 0)^\top, \Sigma_0 = I$
for $t \in \{1, \ldots T\}$ **do**
  Receive an example $\mathbf{x}_t \in \mathbb{R}^d$
  Make prediction: $\hat{y}_t = \text{sgn}\left(\boldsymbol{\mu}_{t-1} \cdot \mathbf{x}_t\right)$
  Receive true label y_t
  suffer loss $\ell^\phi\left(\mathcal{N}\left(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1}\right); (\mathbf{x}_t, y_t)\right)$
  if $\ell^\phi\left(\mathcal{N}\left(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1}\right); (\mathbf{x}_t, y_t)\right) > 0$ **then**
    $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \alpha_t y_t \Sigma_t \mathbf{x}_t, \Sigma_{t+1} = \Sigma_t - \beta_t \Sigma_t \mathbf{x}_t^T \mathbf{x}_t \Sigma_t$
     where α_t and β_t are computed by either
    Proposition 1 (SCW − I)or **Proposition2** (SCW − II);
  end if
end for

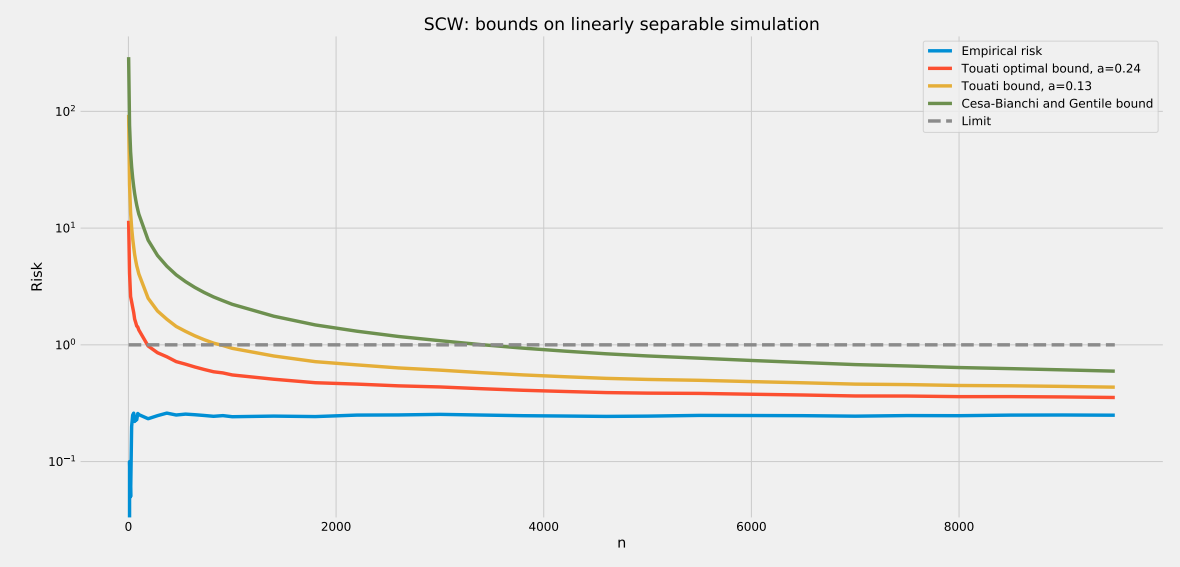


Figure 3.9 – SCW bounds on linearly separable simulation

The behavior of the bound is more or less the same, the only difference is that SCW outperforms PA and therefore we have a relatively smaller impediment risk which improves the efficiency of the bounds.

— Cesa-Bianchi and Gentile (2008) efficient starting from **3700**,minimum=**0.85**.
— Touati optimal bound efficient starting from **400**,minimum=**0.5**.

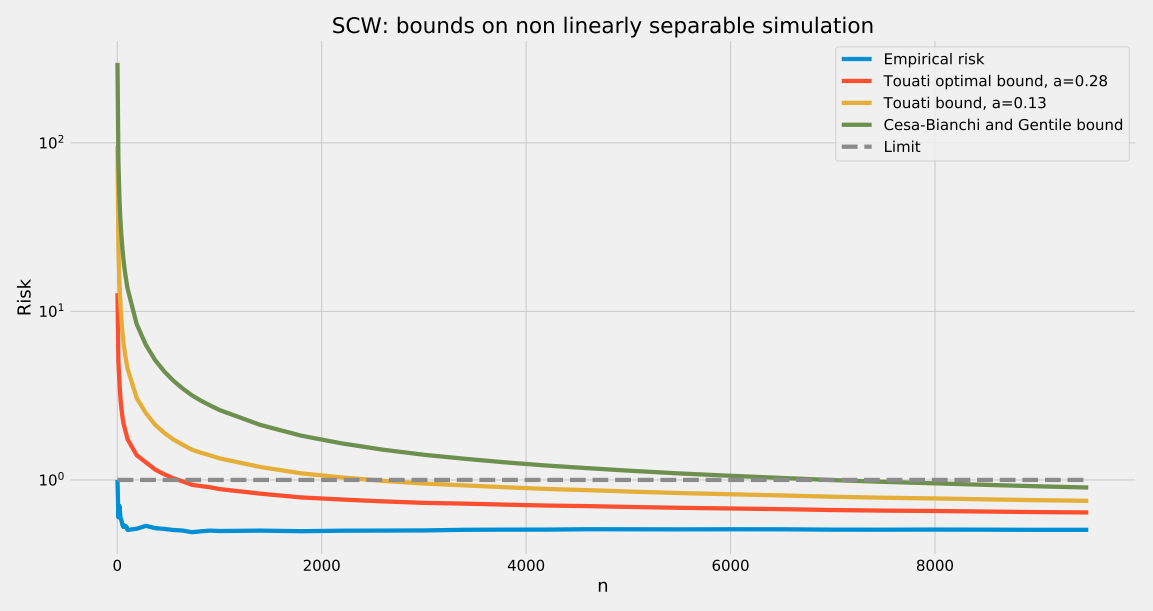


Figure 3.10 – SCW bounds on non linearly separable simualtion

We have the same structure of the plot, however since SCW is not adapted to the non-linearly separable problem, the entangled risk is close to 0.5 and therefore the efficiency of the bounds is deteriorated.

— Cesa-Bianchi and Gentile (2008) efficient starting from **7100**,minimum=**0.95**.
— Touati optimal bound efficient starting from **600**,minimum=**0.63**.

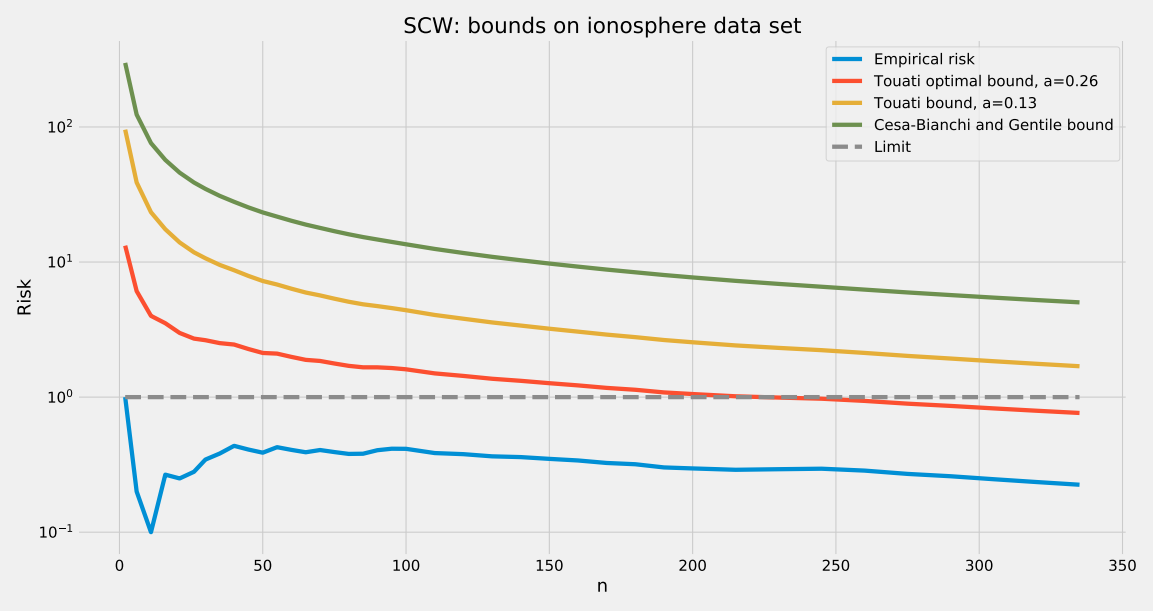


Figure 3.11 – SCW bounds on ionosphere data set

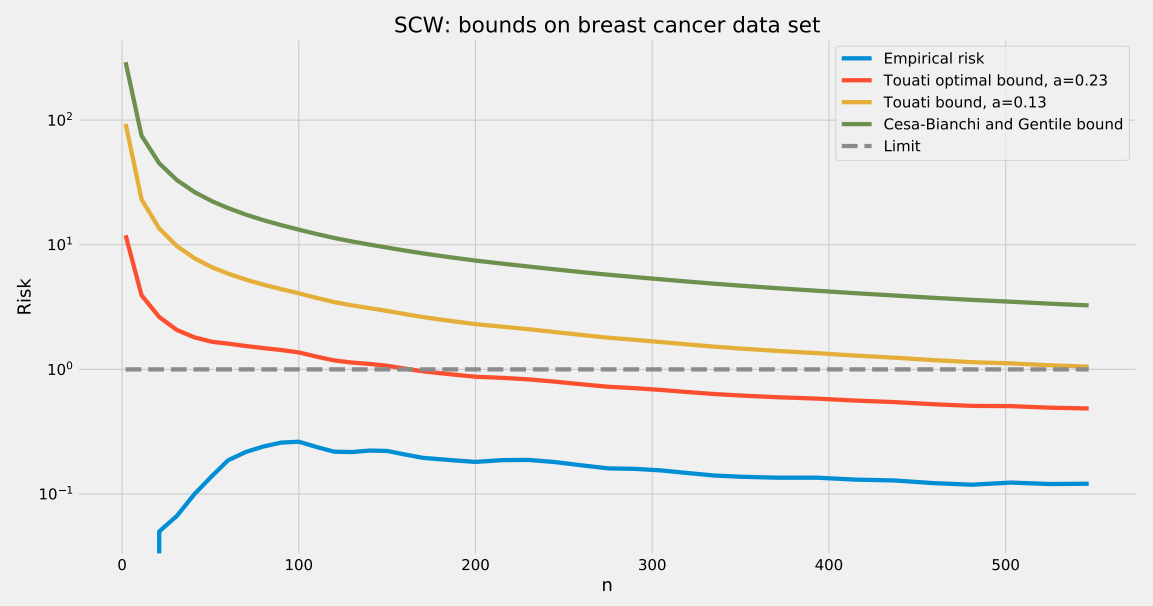


Figure 3.12 – SCW bounds on breast cancer data set

The last two plots confirm once again that, faced with a small sample size and a relatively large empirical risk, our bounds are not informative. Again, those of Cesa-Bianchi and Gentile (2008) are not at all effective because they are always greater than 1.

3.4.4 Multiple layer online perceptron

The Multiple layer Online perceptron is based on the online perceptron the the method that initiated machine learning and artificial intelligence. The pioneering work of Agmon (1954),Rosenblatt (1958) and Novikoff (1963) lay the theoritical foundations of this algorithm. since this is a very well-known reference method in the scientific community, we are not going to offer a detailed summary of it. We recommend a very educational reference Sathyanarayana (2014) to the readers to familiarize themselves with the notion of backpropagation and the architecture of neural networks, notamment the number of layers.

We notice on the one hand that for all the situations, our bounds are drasticly tighter than those of Cesa-Bianchi and Gentile (2008) for any a even in the neighborhood of $\frac{1}{8}$, on the other hand, our bounds converge faster towards the minimum value of the empirical risk. We specify that this algorithm enjoys very high precision. We are evaluating the bounds for very small values of the empirical risk.

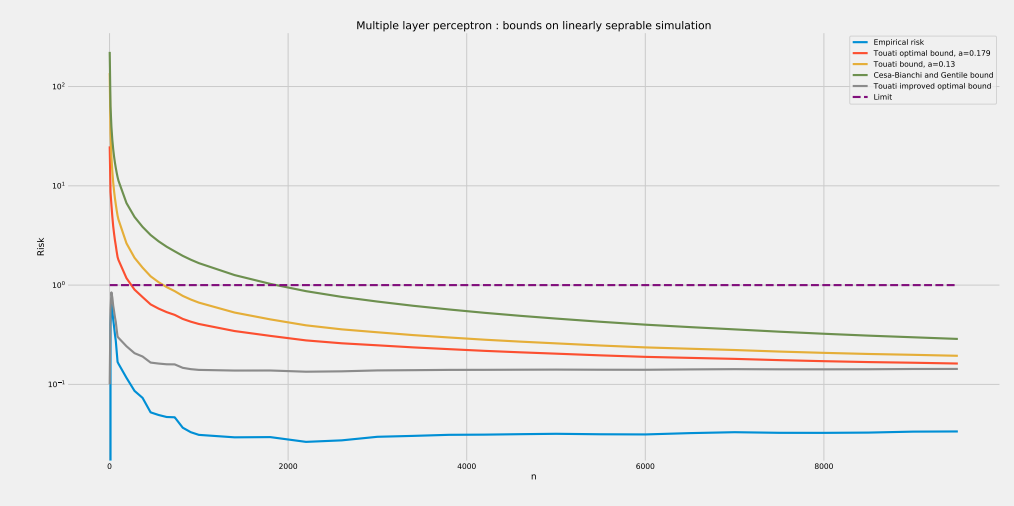


Figure 3.13 – MLP bounds on linearly separable simulation

The Cesa-Bianchi and Gentile (2008)bound is effective from a sample size equal to **1870**, it reaches its minimum value **0.28**, Touati optimal bound is effective from a **250** sample size, it reaches its minimum value **0.16**. Even for a very large sample size, our bounds are almost twice as accurate as those of Cesa-Bianchi and Gentile (2008). The improved optimal bound is always effective, it reaches its minimum value **0.14**. For very large n, the difference becomes blurred between Touati optimal bound and the improved optimal bound.

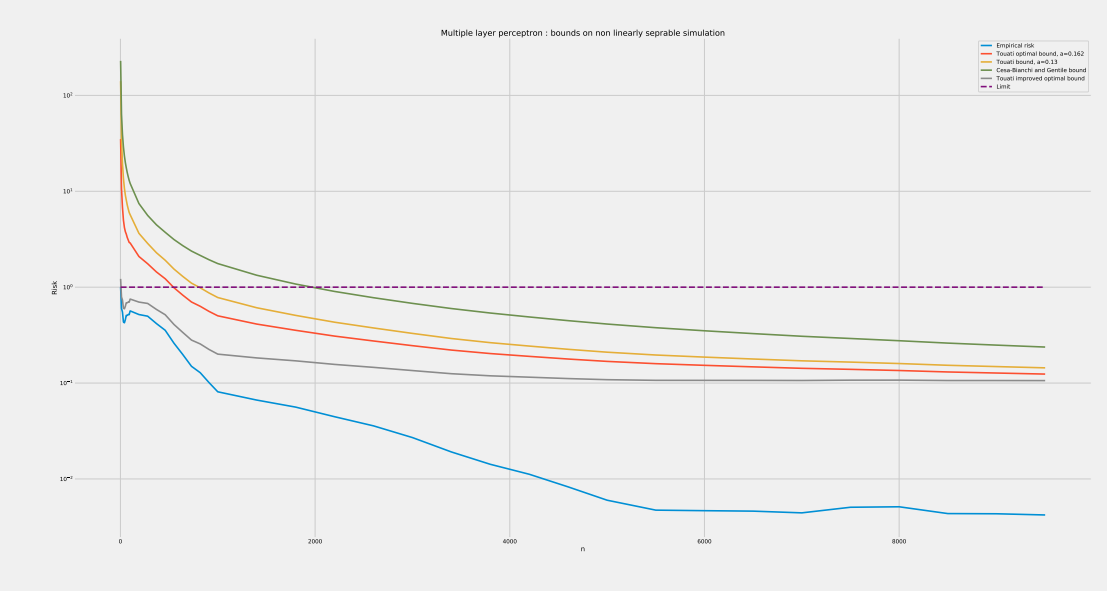


Figure 3.14 – MLP bounds on non linearly separable simulation

In this example, MLP is very suitable for this binary classification and perfom a near faultless. The behavior of the terminals is almost the same:

— Cesa-Bianchi and Gentile (2008) Bound efficient starting from **1990**, minimum=**0.24**.
— Touati optimal Bound efficient starting from **,** minimum=**0.12**.
— The improved optimal bound is always effective, minimum=**0.1**.

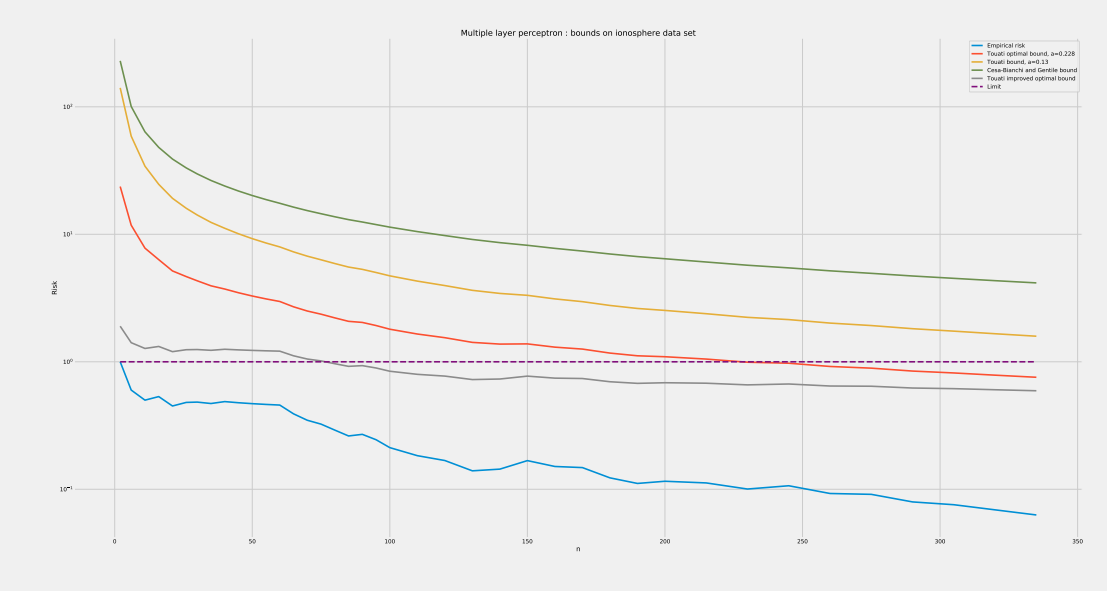


Figure 3.15 – MLP bounds on inonosphere set

Cesa-Bianchi and Gentile (2008) Bound is never efficient in this example which confirms the findings in previous simulations for SCW and PA.

— Touati optimal Bound efficient starting from **230** minimum=**0.75**.

— The improved optimal is efficient strating from **75**, minimum=**0.6**.

We notice that when the sample size is very small and faced with an empirical risk outside a neighborhood of 0, our bounds are not very informative insofar as we know with certainty that the empirical risk is smaller than **0.5**.

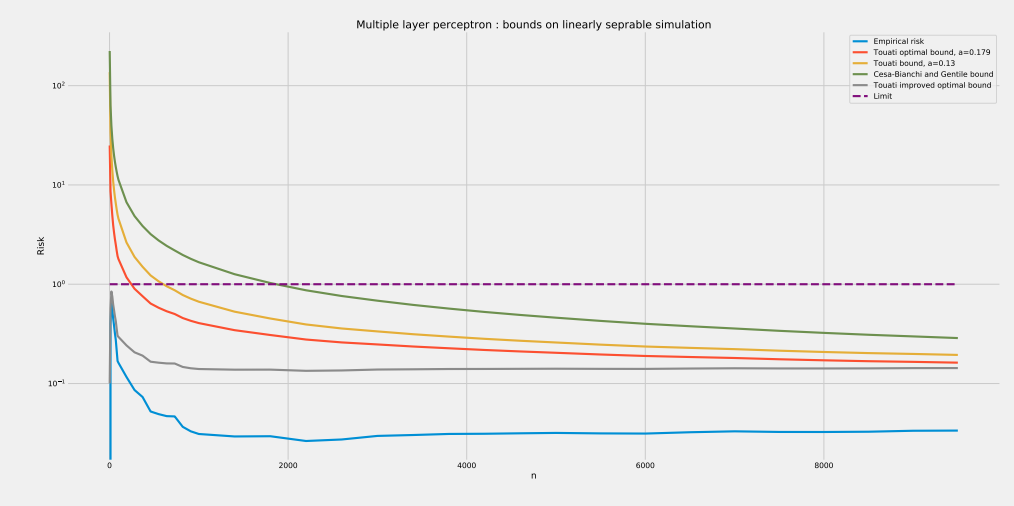


Figure 3.16 – MLP bounds on breast cancer data set

— Touati optimal Bound efficient starting from **245** minimum=**0.65**.
— The improved optimal is efficient strating from **35**, minimum=**0.5**.
Same conclusion.

3.5 Proofs of the main results

3.5.1 Proof of Lemma 3.1

Proof. By keeping the same notations presented beforehand:
we set $\widehat{T} = \text{argmin}_{0\leqslant t<n}\{\text{PER}_{n,\delta}(h_t)\}$ and $T^\star = \text{argmin}_{0\leqslant t<n}\{R(h_t)+F_\delta(R(h_t),t)\}$. We further set $h^\star = h_{T^\star}$ and $r^\star = r_{T^\star} = \widehat{R}_n^{(T^\star)}$. By construction one has $\widehat{h} = h_{\widehat{T}}$. Hence we have for $A^\star = A_{T^\star}$ to be fixed later,

$$\begin{aligned}
&\mathbb{P}[R(\widehat{h}) > R(h^\star) + F_\delta(R(h^\star), T^\star)] \\
&\qquad = \mathbb{P}[\{R(\widehat{h}) > R(h^\star) + F_\delta(R(h^\star), T^\star)\} \cap \{R(h^\star) \geqslant r^\star - A_{T^\star}\}] \\
&\qquad\quad + \mathbb{P}[\{R(\widehat{h}) > R(h^\star) + F_\delta(R(h^\star), T^\star)\} \cap \{R(h^\star) < r^\star - A_{T^\star}\}] \\
&\qquad \leqslant \mathbb{P}[R(\widehat{h}) > R(h^\star) + F_\delta(r^\star - A_{T^\star}, T^\star)] + \mathbb{P}[R(h^\star) < r^\star - A_{T^\star}]
\end{aligned}$$

where the last inequality is due to $r \to F_\delta(r,\cdot)$ is incresing in r. Now the event $\{R(h^\star) < r^\star - A_{T^\star}\} \subset \cup_{t=0}^{n-1}\{R(h_t) < \widehat{R}_n^{(t)} - A_t\}$. Then we get

$$\begin{aligned}
&\mathbb{P}[R(\widehat{h}) > R(h^\star) + F_\delta(\widehat{R}(h_{T^\star}), T^\star)] \\
&\qquad \leqslant \mathbb{P}[R(\widehat{h}) > R(h^\star) + F_\delta(r^\star - A_{T^\star}, T^\star)] + \sum_{t=0}^{n-1} \mathbb{P}[R(h_t) < \widehat{R}_n^{(t)} - A_t].
\end{aligned}$$

To control the term $\mathbb{P}[R(h_t) < \widehat{R}_n^{(t)} - A_t]$, we use (**Theorem 2.28 page 31**) in Bercu et al. (2015),

$$\begin{aligned}
\mathbb{P}[R(h_t) < \widehat{R}_n^{(t)} - A_t] &= \mathbb{P}[\widehat{R}_n^{(t)} - R(h_t) > A_t] \\
&\leqslant \exp\left(-(n-t)V_n.A_t^2 h(\frac{A_t}{V_n})\right)
\end{aligned}$$

where

$$\begin{aligned}
V_n &= \frac{1}{n-t}\sum_{i=t+1}^{n} \mathrm{E}\left[\ell^2\left(H_{i-1}(X), Y\right)\right] \\
\text{and } \forall x \in [0,1]\ h(x) &= (x+1).ln(x+1) - x
\end{aligned}$$

Since $V_n \leqslant R(h_t)$ we obtain:

$$\begin{aligned}
\mathbb{P}[R(h_t) < \widehat{R}_n^{(t)} - A_t] &= \mathbb{P}[\widehat{R}_n^{(t)} - R(h_t) > A_t] \\
&\leqslant \exp\left(-(n-t)R(h_t).A_t^2 h(\frac{A_t}{R(h_t)})\right)
\end{aligned}$$

Let $B = (n-t)R(h_t).A_t^2 h(\frac{A_t}{R(h_t)})$ then

$$A_t = R(h_t).h^{-1}\left(\frac{B}{(n-t).R(h_t)}\right)$$

The function h is not analytically invertible, we lower bound it by an invertible one. Since $\forall x \in [0,1]\ h(x) \geqslant \frac{5}{13}.x^2$ thus $h^{-1}(x) \leqslant \sqrt{\frac{13}{5}}.x$.

Therefore

$$A_t \leqslant \Psi_B(R(h_t), t)$$

Therefore, we arrive at

$$\sum_{t=0}^{n-1} \mathbb{P}[R(h_t) < \widehat{R}_n^{(t)} - \Psi_B(R(h_t), t)] \leqslant n.\exp(-B).$$

The expression $\mathbb{P}[R(\widehat{h}) > R(h^\star) + F_\delta(r^\star - A_{T^\star}, T^\star)]$ is upper bounded, using the same techniques of inequalities **(6)** and **(7)** of Cesa-Bianchi and Gentile (2008) by:

$$n.e^{-B} + n.\sum_{t=0}^{n-1} \mathbb{P}[R(h_t) - \widehat{R}_n^{(t)} > pen_\delta(R(h_t), t)]$$

if we apply **Theorem 1.6**

$$\sum_{t=0}^{n-1} \mathbb{P}[R(h_t) - \widehat{R}_n^{(t)} > pen_\delta(R(h_t), t)] = n.\exp(-B)$$

if and only if

$$pen_\delta(R(h_t), t) = \sqrt{\frac{(1-R(h_t))^2}{\log(\frac{1}{R(h_t)})}.\frac{B}{n-t}}$$

In final

$$\mathbb{P}[R(\widehat{h}) \leqslant \min_{0\leqslant t<n}\{R(h_t) + F_\delta(R(h_t), t)\}] \leqslant (n^2 + 2.n).\exp(-B)$$

If $B = \frac{n.(n+2)}{\delta}$, we get the expected result. □

3.5.2 Proof of Theorem 2.1

Proof. Let $R_n^{(t)} = \frac{1}{n-t}\sum_{i=t}^{n-1} R(h_i)$, We apply the lemma 2.1 with $\text{pen}_{\frac{\delta}{2}}(.,.)$. We obtain

$$\mathbb{P}\left[R(\widehat{h}) > \min_{0\leqslant t<n} R(h_t) + F_{\frac{\delta}{2}}(R(h_t), t)\right] \leqslant \frac{\delta}{2} \tag{3.2}$$

Since

$$\begin{aligned} E = & \min_{0\leqslant t<n} R(h_t) + F_{\frac{\delta}{2}}(R(h_t), t) \\ = & \min_{0\leqslant t<n} \min_{t\leqslant i<n} R(h_i) + pen_{\frac{\delta}{2}}(R(h_i) + \Psi_C(R(h_i), i), i) \\ \leqslant & \min_{0\leqslant t<n} \frac{1}{n-t}\sum_{i=t}^{n-1} R(h_i) + pen_{\frac{\delta}{2}}(R(h_i) + \Psi_C(R(h_i), i), i) \end{aligned}$$

Using **Lemma 1.10**:

$$(E) \leqslant R_n^{(t)} + \frac{\sqrt{\frac{C}{3}}}{n-t}\sum_{i=t}^{n-1}\sqrt{1+4.(R(h_i)+\Psi_C(R(h_i),i))+(R(h_i)+\Psi_C(R(h_i),i))^2}\frac{1}{\sqrt{n-i}}$$

Thanks to Cauchy-Schwarz inequality and integral test for convergence we obtain :

$$\begin{aligned}(E) &\leqslant \min_{0\leqslant t<n} R_n^{(t)} + \sqrt{\frac{C}{3}}\Big(\sqrt{1+4.(R_n^{(t)}+\Psi_C(R_n^{(t)},t))+(R_n^{(t)}+\Psi_C(R_n^{(t)},t))^2}\Big)\sqrt{\frac{\log(n-t)}{n-t}} \\ &= \min_{0\leqslant t<n} g_t(R_n^{(t)}) \qquad (3.3)\end{aligned}$$

It is obvious that $M_n^{(t)} = R_n^{(t)} - \widehat{R}_n^{(t)}$ is a locally squared real Martingale.
(i.e $M_n^{(t)} = M_n - M_{n\wedge t}$)
By applying **Inequality 2.40** with $\frac{\delta}{2n}$, we obtain for
$\forall n \geqslant am(a)\log(\frac{2n}{\delta})$

$$\mathbb{P}\Big(M_n^{(t)} \geqslant \frac{ac(a)\log(\frac{2n}{\delta})}{n-t} + \sqrt{\frac{a\Delta_n(a)\log(\frac{2.n}{\delta})}{n-t}}\Big) \leqslant \delta$$

$$\Delta_n(a) = 2 + 2c(a)\widehat{R}_n^t + ac^2(a)\log(2.n/\delta)/(n-t)$$

. ($c(a)$ and $m(a)$ are defined in the previous chapter.) This expression is equivalent to

$$\mathbb{P}\Big(R_n \geqslant \Phi_a^{-1}(\widehat{R}_n^{(t)})\Big) \leqslant \frac{\delta}{2n}, \quad \forall \delta \in]0,1]. \qquad (3.4)$$

n being fixed beforehand.

We introduce now, the random variables $Z_0, ...Z_{n-1}$ with
$Z_t = g_t(\Phi_a^{-1}(\widehat{R}_n^{(t)}))\forall t \in \{0, ..., n-1\}$.
Thus we can write:

$$\begin{aligned}&\mathbb{P}\Big(\min_{0\leqslant t<n}(R(h_t)+pen_{\frac{\delta}{2}}(R(h_t)+\Psi_C(R(h_t),t),t)) \geqslant \min_{0\leqslant t<n} Z_t\Big) \\ &\leqslant \mathbb{P}\Big(\min_{0\leqslant t<n} g_t(R_n^{(t)}) \geqslant \min_{0\leqslant t<n} g_t(\Phi_a(R_n^{(t)}))\Big) \\ \leqslant &\sum_{t=0}^{n-1}\mathbb{P}\Big(g_t(R_n^{(t)}) \geqslant g_t(\Phi_a(\widehat{R}_n^{(t)}))\Big)\end{aligned}$$

by monotonicity of $g_0, ..., g_{n-1}$

$$\leqslant \sum_{t=0}^{n-1}\mathbb{P}\Big(R_n^{(t)} \geqslant \Phi_a^{-1}(\widehat{R}_n^{(t)})\Big) \leqslant \frac{\delta}{2}$$

by referring to (3.4), combining with (3.2) complete the proof.

□

Conclusion

3.6 Summary and main contributions

This work was intended to develop mathematical tools address to concentration inequalities that aim to deal with martingales. The two main themes covered in this manuscript are exponential inequalities for self-normalized martingales and risk tail bounds for online learning algorithms. The first theme is covered in Chapter 2, the second is treated in Chapter 3. A state of the art of existing methods and mathematical tools necessary for the whole thesis is provided in Chapter 1. At first we have presented, classical inequalities for the sums of bounded random variables of type Bernstein (1927) and Bennett (1962) and their improvements. We then introduced exponential inequalities for martingales, and finally we have briefly presented the field of online learning, its theoretical foundations and the close link it has with concentration inequalities.

Chapter 2 aims to establish new concentration inequalities for self-normalized martingales and set up some elementary applications in statistics and online learning.
It is mainly a scientific contribution in probability theory which on the one hand, allows to standardize the previous inequalities of the literature and on the other hand guarantees the possibility of reaching an optimality of the bound for each application.
This work has been promoted in the form of a mathematical article published in **Electronic Communications in Probability** Bercu and Touati (2019), since summer 2019.
We had the chance to present a summary of it in the form of a talk at the well-known: **Ecole d'été de Saint-Flour**.

Chapter 3 is direcly linked to the previous one. Thanks to the probabilistic tools developed throughout Chapter 2, we develop much more precise risk tail bounds than those of reference in the field, namely those of Cesa-Bianchi and Gentile (2008), world-renowned scientists in the field. What is interesting in our approach; is that with these same probabilistic tools we can improve a whole panoply of results in the field of online learning, thus opening up a field of research to be explored. The synthesis of this work is being submitted to **Journal of Machine Learning Research**.

3.7 Perspectives and future directions

The results in this thesis lead to some future directions as follows:

— Extend our results on autoregressive process with non-symmetric innovations to autoregressive process with ARCH innovations.It would be wise to compare yourself to reference works like that of Claudia Klüppelberg.

— From a probabilistic point of view, we can think of extending our inequalities of concentrations to matrix martingales. Tropp (2011) did so for Freedman inequalities. This direction is interesting from a theoretical point of view because we are going to handle another more complex mathematical object but also stimulating for applications in particular for Multi-armed bandit problems.

Regarding the perspectives we are working with Prof. Odalric Maillard on improving the work of Neu (2015) and Lee et al. (2020) concerning high-probability data-dependent regret bounds for adversarial bandits. The second part of this work consists in the elaboration of structural inequalities based on Bercu and Touati (2019) in order to apply them to the aggregation of experts for unbounded convex losses.

List of Figures

Bibliography

Agmon, S. (1954). The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:382–392. 77

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367. 20, 22, 29, 59

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45. 19, 83

Bercu, B. and Chafaï, D. (2007). *Modélisation stochastique et simulation-Cours et applications*. Dunod. 14

Bercu, B., Delyon, B., and Rio, E. (2015). *Concentration inequalities for sums and martingales*. Springer. 7, 8, 24, 30, 34, 35, 43, 80

Bercu, B., Gamboa, F., and Rouault, A. (1997). Large deviations for quadratic forms of stationary gaussian processes. *Stochastic Processes and their Applications*, 71(1):75–90. 33

Bercu, B. and Touati, A. (2008). Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5):1848–1869. 24, 30, 33, 42, 55

Bercu, B. and Touati, T. (2019). New insights on concentration inequalities for self-normalized martingales. *Electronic Communications in Probability*, 24(63). 8, 29, 54, 59, 60, 63, 83, 84

Bernstein, S. (1927). Probability theory, moscow. *GOS. Publishing house*. 15, 83

Bernstein, S. (1937). Sur quelques modifications de l'inégalité de tchebycheff. In *CR (Doklady) Acad. Sci. URSS*, volume 17, pages 279–282. 23

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer. 51

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press. 30

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057. 48, 49, 60

Cesa-Bianchi, N. and Gentile, C. (2008). Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390. 14, 48, 49, 51, 54, 59, 60, 61, 63, 64, 68, 69, 74, 75, 76, 77, 78, 79, 81, 83

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press. 25, 48

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585. 66

Crammer, K., Kulesza, A., and Dredze, M. (2009). Adaptive regularization of weight vectors. In *Advances in neural information processing systems*, pages 414–422. 70

Crépel, P. (1984). Quelques matériaux pour l'histoire de la théorie des martingales (1920-1940). *Publications mathématiques et informatique de Rennes*, pages 1–66. 11

Dacunha-Castelle, D. and Duflo, M. (1985). *Probability and Statistics: Volume II*, volume 2. Springer Science & Business Media. 14

De la Peña, V. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564. 23, 24, 34

De la Peña, V., Klass, M. J., and Lai, T. L. (2007). Pseudo-maximization and self-normalized processes. *Probability Surveys*, 4:172–192. 30

De la Peña, V. and Pang, G. (2009). Exponential inequalities for self-normalized processes with applications. *Electronic Communications in Probability*, 14:372–381. 30, 57

Delyon, B. (2009). Exponential inequalities for sums of weakly dependent variables. *Electronic Journal of Probability*, 14:752–779. 24, 30, 42, 48, 55, 57

Diaconis, P. and Fulton, W. (1991). A growth model, a game, an algebra, lagrange inversion, and characteristic classes. *Rend. Sem. Mat. Univ. Pol. Torino*, 49(1):95–119. 42, 43

Dredze, M., Crammer, K., and Pereira, F. (2008). Confidence-weighted linear classification. In *Proceedings of the 25th international conference on Machine learning*, pages 264–271. 70, 71

Duflo, M. (1997). *Random iterative models*, volume 34. Springer Science & Business Media. 14

Fan, X., Grama, I., and Liu, Q. (2015). Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20. 30

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, 3:100–118. 48, 59

Freund, Y. and Schapire, R. E. (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103. 27

Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic press. 14

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325. 27

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30. 16, 19, 20, 29

Hoi, S. C., Sahoo, D., Lu, J., and Zhao, P. (2018). Online learning: A comprehensive survey. *arXiv preprint arXiv:1802.02871*. 7, 24

Kearns, M. and Saul, L. (1998). Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 311–319. Morgan Kaufmann Publishers Inc. 21

Lawler, G. F., Bramson, M., and Griffeath, D. (1992). Internal diffusion limited aggregation. *The Annals of Probability*, pages 2117–2140. 43

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. (2020). Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 33. 84

Lévy, P. (1937). Théorie de l'addition des variables aléatoires. *Gauthiers-Villars, Paris*. 11

Liptser, R. and Shiryayev, A. N. (1989). *Theory of martingales*, volume 49. Springer Science & Business Media. 14

Locker, B. (2009). Doob at lyon. *Electronic J. History Probab. Statist*, 5. 11

Mazliak, L. (2009). How paul lévy saw jean ville and martingales. *Electronic Journal for History of Probability and Statistics (www. jehps. net)*, 5. 11

Mazliak, L., Priouret, P., and Baldi, P. (1998). *Martingales et chaînes de Markov*. Hermann. 14

Neu, G. (2015). Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3168–3176. 84

Neveu, J. (1972). *Martingales à temps discret*, volume 9. Masson Paris. 14

Novikoff, A. B. (1963). On convergence proofs for perceptrons. Technical report, STANFORD RESEARCH INST MENLO PARK CA. 77

Pinelis, I. (2014). On the bennett-hoeffding inequality. *Annales de l'IHP Probabilités et statistiques*, 50(1):15–27. 30

Rakhlin, A. and Sridharan, K. (2017). On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory*, pages 1704–1722. PMLR. 28

Rio, E. (2013). Extensions of the hoeffding-azuma inequalities. *Electronic Communications in Probability*, 18(54). 30

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386. 77

Sathyanarayana, S. (2014). A gentle introduction to backpropagation. *Numeric Insight*, 7:1–15. 77

Shafer, G. (2009). The education of jean andré ville. *J Electron Hist Probab Stat*, 5(1). 11

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30. 27, 51

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266. 66

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270. 84

van de Geer, S. A. (2002). On hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data*, pages 161–169. Springer. 21, 22

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999. 26, 27

Ville, J. (1939). Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824. 11

Von Mises, R. (1932). Wahrscheinlichkeitsrechnung und ihre anwendung in der statistik und theoretischen physik. *Bull. Amer. Math. Soc*, 38:169–170. 11

Wang, J., Zhao, P., and Hoi, S. C. (2012). Exact soft confidence-weighted learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 107–114. 70, 71, 72, 73

White, J. S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *The Annals of Mathematical Statistics*, pages 1188–1197. 33, 41

Résumé

Au cours des deux dernières décennies, le domaine des inégalités de concentration a connu un essor important aussi bien en probabilités que pour les domaines applicatifs. Cette thèse apporte en premier lieu une mise en perspective exhaustive de la littérature sous-jacente au domaine. Les deux principales contributions scientifiques s'articulent autour de nouvelles inégalités de concentrations pour les martingales auto-normalisées avec des applications en statistiques d'une part, d'autre part un perfectionnement significatif des bornes de risques pour des algorithmes d'apprentissage séquentiel. En outre, nous connectons grâce à cette thèse deux domaines jusque-là assez éloignés, à savoir les nouvelles inégalités de concentrations pour les martingales, les améliorations des inégalités de type Bernstein avec le domaine de l'apprentissage automatique séquentiel.

Mots-clefs:

Inégalités de concentrations, Martingales, Statistiques Apprentissage automatique séquentiel, bornes de risque.

Abstract

The field of concentration inequalities has gained a significant traction over the last two decades, from contributing to the resolution of complex applied problems to enhancing the theoretical framework of probability. The thesis provides a thorough and exhaustive overview of the relevant scientific literature.
The two main components of this scientific contribution are focused on developing new concentration inequalities for self-normalised martingales with applications to statistics, as well as a drastic improvement to the risk tail bounds of online machine learning algorithms. This work succeeded to connect two fields that have been relatively distinct. The findings bridge the gap between the field of online machine learning and the new concentration inequalities for both martingales and the improvements of the Bernstein type inequalities for random variables.

Keywords: Concentrations Inequalities, Martingales, Statistics, Online Machine Learning, risk bounds.