Machine Learning sur les séries temporelles et applications à la prévision des ventes pour l'E-Commerce

Rémy Garnier

École doctorale nº 405 : Economie, Management, Mathématiques, Physique et Sciences Informatiques

Doctorat

THÈSE

pour obtenir le grade de docteur délivré par

CY Cergy Paris Université

Spécialité doctorale "Mathématiques"

présentée et soutenue publiquement par

Rémy GARNIER

le 8 décembre 2021

Machine Learning sur des séries temporelles dépendantes et applications à la prévision des ventes pour l'E-Commerce

Directeur de thèse : **Paul DOUKHAN**
Co-Directeur de thèse : **Joseph RYNKIEWICZ**
Encadrant industriel : **Bruno GOUTORBE**

Jury

M. Pierre Alquier,	RIKEN AIP Tokyo	Examinateur
M. Paul Doukhan,	CY Cergy Paris Université	Directeur de thèse
M. Gilles Fortin-Stoltz,	Université Paris-Saclay, CNRS	Examinateur
M. Bruno Goutorbe,	CDiscount	Examinateur
Mme Madalina Olteanu,	Paris-Dauphine PSL University	Rapporteure
Mme Anne Philippe,	Université de Nantes	Examinatrice
M. Joseph Rynkiewicz,	'Université Paris 1 Panthéon-Sorbonne	Co-Directeur de thèse
M. Lionel Truquet,	ENSAI	Rapporteur

Invités

M. Jean Marc Bardet,	'Université Paris 1 Panthéon-Sorbonne
Mme Karine Bertin,	Université de Valparaiso
M. Yannig Goude,	Universite Paris-Saclay, EDF

Laboratoire AGM UMR 8088
CY Cergy Paris Université
2 Bd.Adolphe Chauvin,95000 Cergy-Pontoise France

THESE

Table des matières

Remerciements

La thèse de doctorat est une expérience complexe, aux facettes multiples. La mienne a donné lieu à de nombreuses rencontres enrichissantes, et m'a énormément appris dans de nombreux domains.

En premier lieu, j'aimerai remercier mes directeurs de thèse, Paul Doukhan et Joseph Rynkiewicz, qui m'ont encouragé et aidé à accomplir mon travail de thèse. Je suis reconnaissant pour leurs conseils, qui, même si je ne les ai pas toujours suivis, m'ont été précieux pour avancer. Je remercie également mes encadrants industriels, Arnaud Belletoile et Bruno Goutorbe qui m'ont aidé à naviguer dans l'entreprise CDiscount, à accéder aux données, et à améliorer ma façon de présenter mes résultats.

Je remercie également les équipes du laboratoire SAMM, qui m'ont accueilli en leur sein et m'ont permis de travailler à Paris et d'assister à leurs conférences . Je remercie bien sur Raphaël Langhendries, avec qui les discussions ont été nombreuses et fructueuses et ont mené à plusieurs des chapitres du présent manuscrit. Je remercie également Pierre Alquier, Karine Bertin et Gilles Fortin-Stoltz avec qui la collaboration a été heureuse, même si les travaux issus de cette collaboration ne figurent pas dans le manuscrit final.

Je remercie particulièrement Madalina Olteanu et Lionel Truquet qui ont accepté de prendre le temps de relire le présent manuscrit de thèse. Je remercie également les autres membres du jury, Anne Philippe, Gilles Fortin-Stoltz et Pierre Alquier. J'ai eu la joie de pouvoir travailler avec certains d'entre eux et je suis honoré qu'ils aient accepté de participer à l'évaluation de cette thèse.

Je tiens à remercier également l'entreprise CDiscount, qui m'a accueilli pendant mon doctorat et a financé ma thèse. J'espère que mon travail pourra leur être utile. Je remercie mes collègues Data Scientists de CDiscount avec qui les discussions ont été intéressantes et fructueuses : Axel Bellec, Victor Lecointre, Thomas Lentali, Lena Kernen, Marco Belvilacqua et tous les autres. Également chez Cdiscount, je remercie Romain Morales, Jacques Mabille, Elsa Natoli, Tanguy Naut, Ghenly Sek, qui m'ont aidé à bien comprendre les enjeux des prévisions, et avec qui j'ai passé de nombreuses heures à discuter de ce qui fait une bonne prévision. Enfin, je remercie les ingénieurs IT qui m'ont aidé à m'améliorer en Python et à industrialiser mon code : Régis Floret , Moustapha Mahfoud , Mathieu Lamarque.

Je remercie aussi chaleureusement Linda Isone, qui a m'aidé à aplanir et à régler les nombreux problèmes administratifs rencontrés au cours de la thèse.

Enfin, je souhaite remercier tous ceux qui, de près ou de loin, m'ont encouragé. Je pense en particulier à Manuel, et à ma famille et mes parents , qui, du surcroit, ont contribué à la relecture de ce manuscrit.

Synthèse

L'apprentissage automatique, ou Machine Learning (ML) est un ensemble de techniques et d'algorithmes qui permettent de faire apprendre à un ordinateur à réaliser une certaine tâche. Ces algorithmes sont parvenus à de très bons résultats pour de nombreux domaines d'applications. Pour certains problèmes leurs performances dépassent celles des humains. C'est le cas pour la reconnaissance d'image, certaines tâches de traitement du langage naturel, ou dans la résolution de certains jeux. Pour chacun de ces domaines, cette résolution s'est la plupart du temps accompagnée de l'émergence d'une classe d'algorithmes dominants, du moins lorsque que les quantités de données en jeu deviennent importantes. Ainsi, l'usage des réseaux de neurones convolutionnels sont prépondérants pour les tâches de traitement d'images et les réseaux récurrents utilisant des mécanismes d'attention pour les tâches de traitement de texte. Cependant, il semble qu'une tâche échappe encore à cette standardisation : il n'existe pas de classe d'algorithmes dominants pour la prévision de séries temporelles, c'est à dire de séries de données évoluant avec le temps.

Prédire les futures valeurs de séries temporelles semble pourtant nécessaire dans de nombreux domaines. On trouve ainsi des applications pour contrôler divers processus industriels, pour modéliser des écosystèmes, des phénomènes physiques ou géologiques, ainsi que dans les domaines de la finance, de l'actuariat et des assurances. Certaines tâches de traitement du langage naturel (NLP) peuvent également être vues comme des tâches de prévision de séries temporelles, notamment la construction de modèle de langage à la base des algorithmes de type *encoder-decoder*. Dans le cadre de cette thèse, on s'intéresse plus précisément à l'application de méthode de séries temporelles pour la prévision des ventes dans le cadre d'une plateforme d'E-commerce.

Définissons formellement le problème de prédiction de séries temporelles. Mathématiquement, une série temporelle est une suite de variables aléatoires (X_t) à valeurs dans un certain espace $\mathscr{X}$ [1]. On dispose d'une certaine réalisation de ces variables aléatoires $(x_1, \ldots, x_n)$, et on cherche à prédire les valeurs les plus probables pour les suivantes. En pratique cela revient à trouver des valeurs $(\widehat{x_{n+1}}, \widehat{x_{n+h}})$ dont on veut qu'elles minimisent :

$$\sum_{i=1}^{h} \mathbb{E}[\mathrm{L}(\widehat{x_{n+i}}, \mathrm{X}_{n+i})]$$

où L est une certaine fonction à valeurs réelles positives telle que $\mathrm{L}(x, y)$ mesure l'erreur commise lorsque l'on prédit x à la place de y. Le nombre entier h appelé *horizon* de prédiction qui est le nombre de valeurs "en avance" que l'on cherche à prédire.

Pour pouvoir prédire le futur d'une série temporelle, on s'appuie en général sur son passé. C'est à dire que $\widehat{x_{t+h}}$ s'exprime à l'aide d'une certaine fonction f_h, dite *fonction d'apprentissage* (ou modèle), que l'on applique aux a dernières valeurs passées connues $\widehat{x_{t+h}} = f_h(x_t, \ldots, x_{t-a})$. On peut donc définir la perte liée à une telle fonction pour un horizon h donné :

$$\mathbb{L}(f_h) = \mathbb{E}[\mathrm{L}(f_h(\mathrm{X}_t, \ldots, \mathrm{X}_{t-a}), \mathrm{X}_{t+h})]\,. \tag{1}$$

1. Ici, on supposera que $\mathscr{X} \subset \mathbb{R}$. On ne considèrera donc pas de séries temporelles dans un espace multidimensionnel. Si une plus grande dimension est nécessaire, on parlera plutôt de séries temporelles multiples.

Pour qu'une telle fonction soit utile, on fait donc la supposition qu'il est possible d'extrapoler le passé . On présume donc qu'il existe une certaine forme de similarité entre les données passées et celles du futur. La manière la plus simple de considérer cette invariance est de supposer que notre série est *stationnaire*, ce qui a pour conséquence qu'il existe une distribution de probabilité $\mathscr{P}$ telle que $X_t \sim \mathscr{P}$, pour chaque t. Dans les résultats théoriques présentés dans cette thèse, on considère donc majoritairement des séries stationnaires. C'est cependant une hypothèse assez forte qui ne se vérifie que rarement, et qui devrait être relâchée pour l'étude de données plus appliquées. En particulier, les notions de saisonnalité et d'effets événementiels jouent un rôle très important en pratique pour la prédiction des ventes, alors que ces notions brisent l'hypothèse de stationnarité.

Il y a deux caractéristiques qui distinguent généralement les problèmes de prévision de séries temporelles stationnaires d'autres problèmes de Machine Learning et qui expliquent en partie les difficultés inhérentes à ces tâches.

D'abord, d'un point de vue théorique, les données d'une même série temporelle présentent des dépendances mutuelles. Cela invalide la plupart des approches du Machine Learning classique, qui reposent sur des distributions supposées indépendantes.

D'autre part, d'un point de vue pratique, pour une série temporelle donnée on a généralement un petit nombre de données relativement à d'autres domaines d'application du Machine Learning. C'est particulièrement le cas dans le cadre de la prédiction des ventes, où le nombre de date observées est très largement inférieur au nombre de produits que l'on considère. Cela n'est pas nécessairement le cas pour des problèmes de ML classiques, où l'on a fréquemment un grand nombre de points de données exploitables.

Ces deux aspects sont abordés dans les deux parties de cette thèse.

Deux cadres pour des données dépendantes

Cadre Markovien et Hold-Out

Commençons par la partie théorique. Il existe de multiples façons de modéliser la dépendance entre données. Une méthode très classique est de considérer que les données sont générées par une chaîne de Markov, c'est à dire qu'il existe une certaine fonction F et des innovations indépendantes (ε_t) telles que :

$$X_t = F(X_{t-a}, \dots, X_{t-1}, \varepsilon_t) \qquad \forall t \in \mathbb{N}$$

On s'intéressera à certaines modélisations de la dépendance autour de la notion de chaîne de Markov dans plusieurs cadres utiles pour le Machine Learning.

Tout d'abord, on étudiera l'utilisation de la méthode de sélection de modèle *hold-out* dans le cadre de séries temporelles dépendantes. Cette méthode très populaire sépare un ensemble d'observations en deux parties. On entraîne les modèles sur la première partie, dite ensemble d'entraînement, puis on choisit le meilleur modèle sur la seconde partie, dite ensemble de validation.

Des résultats de type inégalités oracle et bornes de généralisation ont été obtenus dans le cadre hold-out indépendant [BBL02]. Lorsque l'on a N fonctions d'apprentissage $(f_1, \dots, f_N)$, ces résultats visent à contrôler l'écart $\mathbb{L}(f_{\hat{k}}) - \mathbb{L}(f_{\tilde{k}})$, où $f_{\hat{k}}$ est le modèle que l'on sélectionne et $f_{\tilde{k}}$ le meilleur modèle possible. Dit autrement, on cherche à contrôler que, même si on n'obtient pas le meilleur modèle possible, la performance de celui qui a été sélectionné ne s'en écarte pas trop. Typiquement, dans le cadre indépendant, cette borne s'exprime en $O(\frac{1}{\sqrt{m}})$, où m est le nombre d'éléments de l'ensemble de validation.

Cependant, il n'existait pas à notre connaissance de résultats pour des séries temporelles, c'est-à-dire dans le cas où les ensembles d'entraînement et de validation sont dépendants. En se plaçant dans le cadre de données générées par des chaînes de Markov uniformément ergodiques, avec mes deux co-auteurs (Raphaël Langhendries and Joseph Rynkiewickz) nous avons réussi à établir plusieurs inégalités oracle et des bornes de déviation. D'une part, nous avons montré dans ce cadre un contrôle de l'écart $\mathbb{L}(f_{\hat{k}}) - \mathbb{L}(f_{\tilde{k}})$ en $O(\frac{1}{\sqrt{m}})$. D'autre part, nous avons obtenu, en utilisant

des conditions sur le bruit, des vitesses rapides $O(\frac{1}{m})$ similaires également à celles obtenues dans le cadre indépendant. Ces inégalités restent vraies sans faire d'hypothèse sur les fonctions que l'on cherche à sélectionner, ce qui permet de couvrir un large éventail de méthode de Machine Learning. Il semble donc que l'on puisse utiliser des méthodes de hold-out dans le cadre de chaîne de Markov sans se préoccuper de la dépendance entre ensemble de validation et d'entraînement.

Cadre Non-causal

Un autre cadre envisagé dans cette thèse est la modélisation d'une forme assez complexe de dépendance : le cadre non-causal. Dans ce cadre, la dépendance n'est pas seulement unidirectionnelle, mais peut venir de plusieurs directions simultanément. Pour des séries indexées par les entiers, cela revient à dire que les valeurs présentes dépendent à la fois du passé et du futur. Formellement, cela revient à considérer que la série (X_t) est solution d'une équation de la forme :

$$X_t = F(X_{t-a}, \dots, X_{t-1}, X_{t+1}, \cdot, X_{t+a}, \varepsilon_t), \qquad \forall t \in \mathbb{Z} \tag{2}$$

où (ε_t) est une série d'innovations indépendantes. On étend ainsi le cadre des chaînes de Markov, et on peut même ajouter des dimensions, en considérant des champs aléatoires de vecteurs non-causaux au lieu de séries temporelles. Ce type de construction non-causale a été initialement proposé par [DT07].

Ce cadre a un intérêt dans plusieurs cas. D'une part il permet d'analyser les performances de certains algorithmes dits bidirectionnels très populaires notamment pour les applications en NLP. Ensuite, on peut également s'en servir dans le cadre de données longitudinales, qui ne présentent pas nécessairement une direction temporelle. Ainsi, les pixels d'une image, une densité de population spatiale sont des données qui présentent naturellement une telle structure. Il est également possible de considérer les données longitudinales d'un même problème de prédiction des ventes, c'est à dire les ventes de différents produits à une même date, comme des données non causales.

À partir de ce cadre non causal, dans un travail joint avec Raphaël Langhendries nous prouvons des inégalités de concentration du type Hoeffding en utilisant différentes hypothèses de contraction. Ces inégalités améliorent les résultats asymptotiques de [CW16]. Elles s'obtiennent à l'aide d'une approximation d'un processus non-causal par des processus causaux.

À partir ces inégalités, il est possible d'obtenir des inégalités oracles comme dans le cadre dépendant causal présenté plus tôt. Le contrôle de l'écart pour la sélection de modèle $\mathbb{L}(f_{\hat{k}}) - \mathbb{L}(f_{\tilde{k}})$ se fait alors en $O(\sqrt{\frac{\ln n}{n}})$, en utilisant une condition de contraction sur la fonction F, ce qui est presque aussi bien que dans le cadre dépendant causal.

Prévision de séries temporelles multiples et application au cadre de la prévision des ventes

L'un des problèmes majeurs de la prévision des ventes en tant que problème de Machine Learning est le suivant : de nombreux phénomènes interagissent et contribuent à former les ventes de différents produits. Il est pourtant nécessaire de tenir compte de la plupart de ces effets pour obtenir des prédictions effectives. Parmi ces phénomènes, on peut citer de manière non-exhaustive :

— Des saisonnalités, qui affectent les ventes de produits à caractère cyclique, comme les produits de jardinage ou les jouets.

— Des effets calendaires, liés à certains évènements spécifiques ponctuels (évènements sportifs, soldes, vacances scolaires,...)

— Des promotions, mises en avant et autres effets d'animation commerciale, qui peuvent concerner tout ou partie du catalogue (French Days, journées centrées sur un magasin comme le bricolage, le jardins, ou certaines catégories de produits...).

— Des effets de substitution, certains produits en remplaçant partiellement d'autres. C'est en particulier le cas pour les produits avec un fort renouvellement technologiques comme les ordinateurs ou les téléviseurs.
— Des effets de cannibalisation, les ventes de certains produits pouvant être redirigées vers d'autres à la suite d'une nouvelle mise sur le marché ou de variations ponctuelles de prix.
— Des ruptures de stock ou d'autres problèmes divers qui peuvent rendre un produit temporairement indisponible pour les clients.
— Des effets de tendance, liés à l'activité du site et à des changements relatifs de positionnement commercial dans différentes catégories de produits. Typiquement, un E-commerçant peut décider de se spécialiser dans une catégorie de produits, puis d'agrandir ou de modifier son offre.
— Des perturbations liées à la pandémie Covid 19, qui ont transformé de manière radicale les patterns habituels de ventes dans le E-Commerce.

Ainsi, pour une variation donnée de ventes d'un produit, il peut exister de nombreuses explications possibles. Or, on dispose généralement de très peu de points de données pour un produit. En moyenne, le cycle de vente typique d'un produit est de quelques dizaines de semaines, ce qui par exemple est largement insuffisant pour calculer une saisonnalité annuelle à partir des seules ventes de ce produit. Il est donc nécessaire de trouver un moyen de transférer de l'information entre plusieurs produits de la même catégorie. Dit autrement, il faut réussir à partager les informations entre séries temporelles, se servir des informations récoltées sur des produits passés pour étendre les prédictions à de nouveaux produits similaires.

Dans le cadre d'un site de E-Commerce, ce transfert d'informations est simplifié car on a de nombreuses informations externes sur un produit, qui dépassent le cadre de ses seules ventes passées. Ainsi, on dispose d'informations sur ses caractéristiques, son prix, les avis clients, etc.

Dans le cadre de cette partie portant sur les applications, on discutera d'abord de l'objectif de la prévision des ventes, de l'intérêt et des limites du cadre Machine Learning présenté plus haut. En particulier, il est naturel de chercher à savoir ce que peut être une bonne prévision des ventes d'un point de vue applicatif, et si le cadre présenté dans la partie théorique répond à ces demandes.

Ensuite, on présentera deux modèles effectuant une prévision des ventes et qui autorisent un "transfert" d'informations entre différentes séries temporelles.

Tout d'abord, nous présenterons le modèle ARBoost [GB19] développé au cours de cette thèse. Ce modèle régressif se base sur des algorithmes de boosting classiques pour pouvoir exploiter le grand nombre d'informations disponibles. Si ces algorithmes sont bien connus, la manière de les appliquer et en particulier la manière d'intégrer des phénomènes comme les saisonnalités et les différentes variables externes imprévisibles est innovante. Par ailleurs, ce modèle offre un cadre plus général pour travailler avec des séries temporelles multiples et courtes. Ces modèles, pour les données de CDiscount, améliorent les performances par rapport à des modèles basées sur des méthodes classiques de série temporelle .

Ensuite, on essaiera de traiter spécifiquement un des problèmes présentés plus haut qui n'est pas pris en compte par ARBoost : la cannibalisation. Ce phénomène apparaît lorsque que plusieurs produits sont en concurrence, et que l'augmentation de vente de l'un d'entre eux se fait au détriment des autres.

Ce phénomène, pourtant connu des praticiens, n'a en effet jamais été à notre connaissance modélisé par des algorithmes de prédiction. Ce modèle est développé dans l'article d'Annals of Operations Research [Gar21].

Ce modèle peut être vu comme un cas particulier de modèle de comptage multivarié [FSTD20]. Il repose sur la construction d'une variable cachée représentant la *compétitivité*. Cette compétitivité se base sur des variables externes, comme le prix, ou la marge appliquée à différents produits et sert à expliquer les variations de part de marché des différents produits à l'intérieur d'une même catégorie. Ainsi, lorsque la compétitivité d'un produit augmente relativement à celles des autres, sa part de marché augmente dans la famille considérée.

Pour tenter de rendre mesurable cette variable cachée, le modèle entraîne un réseau de neurones identique pour chacun des produits d'une même catégorie. Le fait de prendre le même réseau permet de partager des informations entre différents produits, et de traiter des séries de données courtes comme dans le cas précédent.

Ce modèle de compétitivité a des bonnes performances lorsqu'on l'applique à des données CDiscount. En particulier, il parvient à éviter un écueil classique des algorithmes de prévisions de vente : la tendance à la sous-prédiction.

Première partie

Machine Learning pour des données dépendantes

Chapitre 1

Introduction

1.1 Machine Learning dans le cadre indépendant

Dans cette section, on rappelle les principes généraux qui sous-tendent les résultats classiques de Machine Learning indépendant. Le but n'est absolument pas d'être exhaustif, mais plutôt de placer les résultats obtenus lors de cette thèse dans leur contexte. On se place ici dans un cadre très classique proche de celui, historique, de Vapnik [Vap99], celui d'une régression supervisée, qui est le cas le plus proche de ce que l'on rencontrera lorsque l'on cherchera à prédire des séries temporelles.

On considère deux ensembles $\mathscr{X}$ et $\mathscr{Y}$. On suppose que l'on observe un jeu de n points de données $\mathrm{D}_n = ((x_1, y_1), \ldots, (x_n, y_n))$. Pour chaque point de donnée i, $x_i \in \mathscr{X}$ est un vecteur de caractéristique, et $y_i \in \mathscr{Y}$ est une étiquette. On en parle plus souvent en utilisant les termes anglais de *feature* et *label*. On suppose que ces données ont été engendrées indépendamment à partir d'un certain couple de variables aléatoires (X, Y) de loi jointe $\mathscr{P}$, autrement dit, on suppose qu'il existe une certaine relation constante entre features et labels.

L'objectif central du Machine Learning est d'entraîner une certaine fonction $g : \mathscr{X} \longrightarrow \mathscr{Y}$ tel que $g(\mathrm{X})$ soit un bon estimateur de Y. La fonction g est appelée *fonction d'apprentissage* (ou modèle). Pour estimer la qualité d'une fonction d'apprentissage, on introduit une certaine perte $\mathrm{L} : \mathscr{Y}^2 \longrightarrow \mathbb{R}_+$ qui évalue la qualité d'un tel estimateur g en définissant une erreur théorique associée à cet estimateur :

$$\mathbb{L}(g) = \mathbb{E}[\mathrm{L}(g(\mathrm{X}), \mathrm{Y})] \tag{1.1}$$

.

Lorsque l'on s'intéresse au Machine Learning d'un point de vue théorique, on introduit en général deux notions. La première est la notion d'estimateur de Bayes g^*, qui est le meilleur estimateur possible parmi toutes les fonctions mesurables de $\mathscr{X}$ vers $\mathscr{Y}$, c'est à dire que $g^* = \underset{g}{\operatorname{argmin}}\, \mathbb{L}(g)$. La seconde notion est celle de l'ensemble de modèle S parmi lequel on cherche notre estimateur. On définit donc également le meilleur modèle théorique $g_{\mathrm{S}}^* = \underset{g \in \mathrm{S}}{\operatorname{argmin}}\, \mathbb{L}(g)$ dans cet ensemble S.

Un algorithme de Machine Learning est un choix d'un ensemble S, et d'une méthode pour choisir un certain estimateur $\widehat{g} \in \mathrm{S}$. Pour évaluer la qualité de l'algorithme, il faut contrôler l'excès de risque : $\mathbb{L}(\widehat{g}) - \mathbb{L}(g^*)$. Contrairement à ce qui se passe en statistique classique, on ne cherche donc pas à savoir si l'on parvient à trouver le "bon" modèle, ou un modèle "proche" du bon modèle, mais si l'on trouve un modèle dont l'erreur n'est pas trop loin de celle du meilleure modèle.

On décompose généralement cet excès de risque en introduisant :

— D'une part l'erreur d'approximation $\mathbb{L}(g_{\mathrm{S}}^*) - \mathbb{L}(g^*)$, qui diminue lorsque l'on augmente l'ensemble de modèle S,

— D'autre part l'erreur d'estimation $\mathbb{L}(\widehat{g}) - \mathbb{L}(g_{\mathrm{S}}^*)$, qui augmente en général avec la taille l'ensemble de modèle S.

Il y a donc un compromis à trouver sur la taille de S.

De nombreux résultats existent pour borner l'excès de risque, mais ils dépendent souvent de S d'une part et de la manière de choisir $\widehat{g}$ au sein de S d'autre part. La manière la plus classique de faire ce dernier calcul est de considérer que $\widehat{g}$ est choisi pour minimiser le risque empirique L_n sur le jeu de données dont on dispose, le risque empirique d'une fonction g étant défini de la manière suivante :

$$L_n(g) = \sum_{i=1}^{n} L(g(x_i), y_i) \tag{1.2}$$

Dans ce cadre, $\widehat{g} = \underset{g \in S}{\operatorname{argmin}}\, L_n(g)$. On dit alors que $\widehat{g}$ est un minimiseur du risque empirique et on dispose d'un grand nombre de résultats permettant de contrôler l'excès de risque. Un résultat classique est le suivant (cf. [BBL03]) :

Théorème 1.1. *Soit $\widehat{g}$ minimiseur du risque empirique sur S fini. Si l'on note N le cardinal de S, alors pour tout $\delta > 0$, on a avec une probabilité $> 1 - \delta$:*

$$\mathbb{L}(\widehat{g}) - \mathbb{L}(g^*) \leq 2\sqrt{\frac{\log(N) + log(\frac{2}{\delta})}{2n}}$$

Ce résultat est très général, car on ne fait pas de supposition sur la distribution $\mathscr{P}$ sous-jacente. On obtient une décroissance en $O(\frac{1}{\sqrt{n}})$. Pratiquement, cela signifie que l'on est quasiment garanti d'améliorer les performances de notre algorithme d'un facteur K lorsque l'on multiplie le nombre de données par un facteur K^2.

On ne s'étendra pas ici sur toutes les variantes de ces résultats dans le cadre indépendant, mais il en existe de nombreuses avec les mêmes vitesses. Cependant, notons qu'il est possible d'étendre ces résultats pour nombre infini de modèles, par exemple en utilisant la notion de dimension de Vapnik-Chervonenkis ([BEHW89]).

Ces résultats sont généralement démontrés en utilisant des inégalités de concentration comme celle de Hoeffding qui contrôle la déviation d'une moyenne de variables aléatoires.

Théorème 1.2. HOEFFDING INEQUALITY [HOE63] *Soit $(Z_1, \ldots, Z_n)$ variables aléatoires indépendantes et identiquement distribuées à valeur dans $\mathscr{X}$ et $g : \mathscr{X} \longrightarrow [a, b]$. On note $S_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$ Pour tout $\varepsilon > 0$, on a :*

$$\mathbb{P}[|S_n - \mathbb{E}[S_n]| > \varepsilon] \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$$

Une méthode pour établir des inégalités oracle dans le cadre dépendant sera donc de chercher à établir des inégalités de concentration. C'est ce qu'on fera au chapitre 3, on on montrera une inégalités de type Hoeffding.

1.2 Machine Learning pour des séries temporelles

1.2.1 Risque théorique, Risque empirique

Le cadre d'une série temporelle introduit deux différences majeures avec le cadre i.i.d. (indépendant identiquement distribué) La première est que l'on introduit une dépendance entre les différents points de données. La seconde est que, généralement, l'approche est auto-régressive, les features utilisées pour prédire les prochaines valeurs deviennent les p dernières valeurs précédentes observées. Typiquement, on a alors un jeu de $n + p$ étiquettes $D_n = (y_t)_{t \in [1, n+p]}$ échantillonnées à partir d'une série de variables aléatoires $(Y_t)_{t \in [1, n+p]}$. On pose alors pour chaque $t \in [p+1, n+p]$, le couple $x_t = (y_t, (y_{t-1}, \ldots, y_{t-p}))$ comme point de donnée, avec y_t l'étiquette du t-ème point de données, et $((y_{t-1}, \ldots, y_{t-p}))$ le vecteur de caractéristiques correspondant. De manière similaire $X_t = (Y_t, (Y_{t-1}, \ldots, Y_{t-p}))$ est la variable aléatoire correspondante.

De la même manière que précédemment, on cherche à construire une fonction $g : \mathscr{Y}^p \longrightarrow \mathscr{Y}$, telle que $g((Y_{t-1},\dots,Y_{t-p}))$ soit une bonne prédiction de Y_t pour tout t. On introduit alors une perte $L : \mathscr{Y}^2 \longrightarrow \mathbb{R}_+$ qui mesure la qualité d'une prédiction, et, avec un léger abus de notation, on notera pour un modèle de prédiction g et une réalisation :

$$L(g(X_t)) \coloneqq L\big(g(Y_{t-1},\cdots,Y_{t-p}),Y_t\big). \tag{1.3}$$

Notre série d'étiquettes est souvent définie comme *stationnaire*, c'est à dire que l'on suppose que pour toute fonction f mesurable et tout entier k, $f(X_1,\dots X_t)$ et $f(X_{k+1},\dots X_{t+k})$ ont la même distribution de probabilité . Une conséquence immédiate est qu'il existe une distribution de probabilité $\mathscr{P}$ telle que $X_t \sim \mathscr{P}$, $(\forall\, t)$. Grâce à cette condition, on peut étendre la notion d'erreur théorique définie par (1.2) . Il s'agit de la perte attendue calculée avec la loi stationnaire. Ce critère peut être vu comme la perte empirique asymptotique pour un futur infini.

Définition 1.1. RISQUE THÉORIQUE. *Pour (Z_0) un vecteur de même distribution $(X_t)_{t\in\mathbb{Z}}$ mais indépendant de $(X_t)_{t\in\mathbb{Z}}$. Le risque (ou erreur) théorique d'un modèle g est alors :*

$$\mathbb{L}(g) = \mathbb{E}\big(L(g(Z_0))\big). \tag{1.4}$$

De la même manière, on peut définir une erreur empirique sur l'échantillon que l'on considère

Définition 1.2. RISQUE EMPIRIQUE. *Le risque (ou erreur) empirique d'un modèle g pour un échantillon $D_n = (y_t)_{t\in[1,n+p]}$ est :*

$$\hat{L}_m(g) = \frac{1}{m}\sum_{k=n+1}^{n+m} L(g). \tag{1.5}$$

Par la loi des grands nombres, la perte empirique $\hat{L}_m(\hat{g}_1^n)$ converge, presque sûrement, vers la perte théorique $\mathbb{L}(\hat{g}_1^n)$:

$$\mathbb{L}(\hat{g}_1^n) \overset{p.s.}{=} \lim_{m\to\infty} \hat{L}_m(\hat{g}_1^n).$$

1.2.2 Principe des inégalités oracle

L'objectif principal des chapitres 2 et 3 est d'établir des inégalités oracle pour des séries temporelles (donc dépendantes). Présentons dans cette sous-section les concepts nécessaires à la compréhension des inégalités oracle dans le cadre des séries temporelles, en étendant les notions évoquées en section 1.1.

De la même manière qu'en section 1.1, on introduit l'estimateur de Bayes g^* :

$$g^* = \arg\min_{g\in\mathscr{F}} \mathbb{L}(g(X_t)). \tag{1.6}$$

On suppose que l'on a un ensemble de S de procédures d'apprentissage. On suppose que cet ensemble est fini, c'est à dire que $S = \{\hat{g}_1,\dots,\hat{g}_N\}$. On veut sélectionner le *meilleur* modèle de prédiction. Dans un monde idéal, un oracle bienveillant pourrait dire quel indice $\tilde{k}$ minimise l'erreur théorique :

$$\tilde{k} = \arg\min_{k\in\{1,\dots,N\}} \mathbb{L}\big((\hat{g}_1^n)_k\big).$$

Cependant, tout ce que nous pouvons faire est de choisir l'indice $\hat{k}$ qui minimise l'erreur de validation empirique :

$$\hat{k} = \arg\min_{k\in\{1,\dots,N\}} \hat{L}_m\big((\hat{g}_1^n)_k\big).$$

Une inégalité d'oracle entre les choix optimaux et empiriques $\tilde{k}$ et $\hat{k}$ peut s'écrire :

$$\mathbb{E}\big(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \mathbb{L}(g^*)\big) \le C\left(\mathbb{L}((\hat{g}_1^n)_{\tilde{k}}) - \mathbb{L}(g^*) + \frac{\Upsilon(n)}{n}\right). \tag{1.7}$$

[1] où C est un facteur au moins aussi grand que 1, $\gamma(n)$ est une fonction à croissance lente et $\mathbb{L}(g^*)$ est la meilleure erreur attendue. On appelle $\frac{\gamma(n)}{n}$ vitesse de l'inégalité oracle, et on s'intéresse généralement à sa vitesse asymptotique. Lorsque $\frac{\gamma(n)}{n} \in \omega(\frac{1}{\sqrt{n}})$, on parle de vitesse rapide. Le concept d'inégalité d'oracle a été initialement proposé par Donoho et Johnstone [DJ94], et a donné lieu à de multiples résultats dans le cadre de séries temporelles dont on présentera quelques exemples dans la prochaine sous-section.

1.2.3 Modéliser la dépendance

Comme son nom l'indique le cadre indépendant n'a pas besoin de méthodes pour modéliser la dépendance entre les points de données. En recvanche dans le cadre dépendant, on est obligé de faire une ou plusieurs hypothèses quant à la distribution pour établir des inégalités oracle. De très nombreuses méthodes ont été introduites, et cette thèse en propose d'ailleurs une relativement nouvelle au chapitre 3. On ne pourra donc pas présenter toutes ces méthodes, mais on peut essayer de donner un panorama des méthodes qui seront utilisées dans les chapitres suivants, ainsi que cert.

Tout au long de cette première partie, on considérera que nos séries de données sont générées par des chaîne de Markov, c'est à dire qu'il existe une certaine fonction F et des innovations indépendantes identiquement distribuées (ε_t) de distribution μ_ε telles que :

$$X_t = F(X_{t-h}, \dots, X_{t-1}, \varepsilon_t) \qquad \forall t \in \mathbb{N}. \tag{1.8}$$

Les chaînes de Markov ont l'avantage d'être suffisamment générales pour modéliser un ensemble de processus assez variés à mémoire finie [2]. Il existe de nombreuses méthodes pour contrôler la dépendance d'une chaîne de Markov ou d'autres types de séries temporelles. Présentons en un échantillon :

Conditions de contraction. Les conditions de contraction sont utilisées de manière classique pour les chaînes de Markov et pour d'autres processus stochastiques. L'idée est de supposer qu'il existe un vecteur$\alpha = (\alpha_1, ...\alpha_i)$ tel que l'on ait pour tous couples d'uplets $(x_1, \dots, x_h)$ et $(x'_1, \dots, x'_h)$ et pour toute variables $\varepsilon \sim \mu_\varepsilon$:

$$\mathbb{E}[\|F(x_1, \dots x_h, \varepsilon) - F(x'_1, \dots, x'_h, \varepsilon)\|] \leq \sum_{i=1}^{h} \alpha_i \|x_i - x'_i\|, \tag{1.9}$$

avec $\sum_{i=1}^{h} \alpha_i < 1$. Ici, $\|\cdot\|$ est une norme quelconque : on peut donc imaginer de jouer sur cette norme. De nombreux résultats se basent sur ce type de modèle, on peut notamment citer [DF15; DDF19] qui obtiennent des inégalités de concentration en se servant de ce genre de conditions dans un cadre markovien, respectivement stationnaire et non stationnaire. On peut également citer [DT20] qui les utilise pour des processus similaires incluant des covariables externes. Un exemple d'application satisfaisant ce type de conditions est présenté au chapitre 6. Enfin [ACKR16] q applique ce type de conditions à des processus de Poisson auto-régressifs avec covariables. Ce genre de condition sera également utilisé au chapitre 3 sur des champs de vecteurs non-causaux.

Ce type de conditions présente l'intérêt d'être locales, c'est à dire que l'on contrôle la dépendance à l'aide d'une équation locale, ce qui a pour avantage d'être plus simple à manipuler, et de faire sens pour un grand nombre de processus.

1. Cette forme a été choisie car elle est proche des inégalités oracles prouvées au cours de cette thèse. Il existe de nombreuses formes d'inégalité oracle, selon la procédure de sélection de modèle considérée, selon que les hypothèses probabilistes

2. Il existe des chaines de Markov à mémoire infinie. Cependant, la mémoire des ordinateurs étant rarement infinie, le gain pratique pour des applications en Machine Learning semble limité.

Temps de mélange et ergodicité uniforme. Les notions de temps de mélange et d'ergodicité uniforme sont également utiles pour contrôler le comportement d'une chaîne de Markov. Ces notions seront formellement définies au chapitre 2, mais on peut essayer d'en donner une première interprétation. Une chaîne de Markov finie est uniformément ergodique lorsqu'il y a convergence uniforme de la distribution de la chaîne de Markov vers la distribution stationnaire, et ce quelque soit la distribution initiale. Le temps de mixing t_{mix} permet d'évaluer la vitesse de convergence d'une chaîne de Markov vers sa distribution stationnaire. Le fait que t_{mix} soit fini est équivalent à l'ergodicité uniforme de la chaîne (voir Roberts et Rosenthal [RR04]).

Ces deux notions permettent notamment à [Pau15] d'établir des inégalités de concentration que l'on utilisera pour démontrer des inégalités oracle dans le prochain chapitre, y compris pour des vitesses rapides dans le cadre du hold-out. Il semble possible de pouvoir estimer ces temps de mixing [WK19].

Conditions de mixing. Les conditions de mixing fortes sont un vaste ensemble de conditions pour des séries qui ne présentent pas une structure spécifique comme des chaînes de Markov, mais qui ont malgré tout une forme d'"indépendance asymptotique". Ces conditions ont été initialement introduites par [Ros56] puis utilisées pour des résultats asymptotiques, mais il est possible de s'en servir pour établir des inégalités de concentration ou des oracles. En tenant compte des différents types d'inégalités oracles, les différents types de mixing et l'ensemble des méthodes de sélection de modèles considérés, le nombre de résultats de ce type est très grand. Parmi tous ces résultats, on peut citer [HS14] et [ALW13] qui obtiennent des vitesses rapides pour des minimiseurs du risque empirique respectivement pour des cas de $\alpha-$ et de $\phi-$mixing , et [MR10] qui obtient des inégalités de concentration sous des conditions de $\beta-$ et $\phi-$ mixing et les applique dans un cadre assez similaire à celui du chapitre 2.

Dépendance faible. La notion de dépendance faible, présentée notamment dans [DDL+07], relâche certaines contraintes des notions de mixing. Cette notion permet de considérer des séries plus générales que des chaînes de Markov. [AW+12b] s'en sert pour proposer des inégalités oracles pour une procédure de sélection en deux étapes. Les auteurs obtiennent des vitesses quasiment semblables au cas i.i.d. présenté en section 1.1.

Notons que ce ne sont pas les seuls types de résultats de type inégalité oracle. Ainsi [CBL06] s'intéresse à la prévision de suites individuelles en ne faisant aucune hypothèse stochastique et en introduisant une notion de regret analogue à la notion de risque théorique que nous allons étudier.

1.3 Contributions

Présentons les contributions des deux prochains chapitres.

1.3.1 Contribution du chapitre 2

Dans le chapitre 2, on s'intéresse à la procédure de validation hold-out. Cette méthode très populaire sépare un ensemble d'observations en deux parties. On entraîne les modèles sur la première partie, dite ensemble d'entraînement, puis on choisit le meilleur modèle sur la seconde partie, dite ensemble de validation. Dans la littérature économétrique, la méthode du hold-out est appelée validation hors échantillon et est la principale méthode pour sélectionner un bon modèle de série chronologique.

Les estimations de hold-out sont bien étudiées lorsque les données sont indépendantes, mais, à notre connaissance, ce n'est pas le cas lorsque l'ensemble de validation n'est pas indépendant de l'ensemble d'apprentissage. Dans ce chapitre, en supposant que nos données suivent une chaine de Markov uniformément ergodique, on énoncera 3 bornes de généralisation et des inégalités d'oracle pour la validation hold-out :

— Une première inégalité oracle de type Hoeffding (Théorème 2.4) basée sur une inégalité de type Hoeffding, ne reposant que sur l'hypothèse d'ergodicité de la chaine de Markov. Cette borne est du même ordre et de même vitesse asymptotique que le cas indépendant.

— Une quasi-inégalité oracle (Théorème 2.7) basée sur une inégalité de type Bernstein. Cette borne est une borne de type "vitesse rapide". L'inégalité obtenue n'est cependant pas tout à fait une inégalité oracle, car un terme additif vient s'ajouter.

— Pour se débarasser de ce terme additif, il faut ajouter une condition de bruit. On obtient alors une "vraie" inégalité oracle (Corollaire 2.34). On montre donc que la méthode de sélection hors échantillon est adaptative aux conditions de bruit.

1.3.2 Contribution du chapitre 3

Dans le chapitre 3, on étend le cadre markovien présenté dans l'équation. (1.8) pour gérer les données non causales. Dans le cas à une dimension, on suppose que le modèle est solution d'une équation du type :

$$X_t = F(X_{t-s}, \dots, X_{t-1}, X_{t+1}, \dots, X_{t+s}, \varepsilon_t), \qquad \forall t \in \mathbb{Z}$$

avec $(\varepsilon_t)_{t\in\mathbb{Z}}$ innovations i.i.d. On généralise immédiatement cette approche aux champs aléatoires indexés par $\mathbb{Z}^K$, c'est-à-dire le cas multidimensionnel. Dans ce cas, les données sont solution de :

$$X_t = F((X_{t+s})_{s\in\mathscr{B}}, \varepsilon_t), \text{ pour un voisinage } \mathscr{B} \text{ des variables aléatoires i.i.d. } (\varepsilon_t)_{t\in\mathbb{Z}}.$$

Ce type de champ aléatoire a été introduit par [DT07] dans les conditions d'existence et d'unicité d'une telle solution.

On cherche à établir des inégalités de concentration sur ces processus en utilisant des hypothèses réalistes sur les données. En particulier, on veut que nos hypothèses soient raisonnables pour des données textuelles. Les principaux résultats sont des inégalités de type Hoeffding (théorèmes 3.1 et 3.2. Les hypothèses principales de ces inégalités sont des versions non-causales des conditions de contraction utilisées dans [ADF19; DDF19; DF15].
La preuve repose sur une approximation locale pratique d'un champ aléatoire non causal par une fonction de nombre fini de variables aléatoires i.i.d., ce qui se rapproche d'une méthode proposée par [CW18] pour obtenir des résultats asymptotiques non-causaux. On fournit également quelques exemples de tels champs aléatoires, en les comparant à d'autres cadres, tels que les techniques de faible dépendance et de mélange. Enfin, on propose une inégalité oracle pour des champs de vecteurs non causaux de la même forme que la section 1.2.2, inspirée de [Lug02a]. Cet oracle est une conséquence directe de l'inégalité de type Hoeffding.

Chapitre 2

Validation hold-out pour les chaînes de Markov

Ce chapitre est une adaptation d'un article écrit en collaboration avec Joseph Rynkiewicz et Raphaël Langhendries, soumis à la revue ESAIM-PS actuellement en cours de relecture. Ce chapitre est organisé comme suit : tout d'abord, on présentera les motivations pour cette étude, puis, dans la section suivante, on présentera le modèle, les notations et les inégalités de concentration pour une chaîne de Markov uniformément ergodique. Ensuite, on établira d'abord des inégalités exponentielles, les bornes de généralisation et les inégalités d'oracles pour le modèle sélectionné par validation croisée. Ces bornes seront améliorées dans la troisième section sous des hypothèses sur l'erreur théorique. Enfin, on affinera ces bornes dans des conditions de bruit et montrera le résultat principal : la méthode hold-out est toujours adaptative aux conditions de bruit pour une chaîne de Markov uniformément ergodique. Les preuves longues seront laissées dans l'annexe A (en anglais).

2.1 Motivation

L'utilisation d'une mesure de la performance basée sur les erreurs hors échantillon est une méthode répandue pour la sélection de modèles. Cela est également vrai pour les séries chronologiques, et la division des données en un sous-ensemble d'apprentissage et un sous-ensemble de validation d'observations futures est une option mise en œuvre dans la plupart des logiciels statistiques ou d'apprentissage automatique. Dans la littérature de prévision traditionnelle, on peut même dire que l'évaluation hold-out est la procédure d'évaluation standard (voir, par exemple, Tashman [Tas00] sur les procédures hold-out pour évaluer l'exactitude des prédictions des modèles). Dans la communauté de l'apprentissage automatique, Cerqueira et al. [CTM20] compare empiriquement les performances entre procédures hold-out et la validation croisée, et a constaté que les procédures hold-out produisent les estimations les plus précises pour les séries en temps réel. Les auteurs pensent que la principale raison de la performance des procédures hold-out est la préservation de l'ordre temporel des observations. Les méthodes hold-out contiennent essentiellement la dernière partie de la série chronologique pour voir à quel point les prédictions du modèle sont précises. Pour le cas distribué indépendant et identique (i.i.d.), les propriétés théoriques de résistance sont bien connues, et, par exemple, dans le cas de classification, elle s'adapte aux conditions de bruit (voir Blanchard et Massart [BM06]). Cependant, il semble exister peu de résultats théoriques pour la méthode du hold-out pour les données dépendantes. Certaines études évaluent les performances asymptotiques de méthodes apparentées comme la validation croisée (voir Arlot et Celisse [AC10]) dans le contexte de la régression (Burman et Nolan [BN92]). D'autres auteurs étudient des méthodes de sélection de modèles pour les données dépendantes, qui se concentrent davantage sur la pénalisation (voir Alquier et Wintenberger [AW12a]) ou sur des mesures de complexité telles que la complexité de Rademacher (Mohri et Kuznetsov [MK17]) ou sur les bornes de stabilité (Mohri et Rostamizedeh [MR10]). La minimisation empirique du risque a

également été étudiée dans le cadre de chaînes de Markov uniformément ergodiques (voir Bin et al. [BHZ09]). Cependant, de tels résultats sont difficiles à appliquer à des modèles massifs comme les réseaux profonds (Zhang et al. [ZBH+21]), et les procédures hold-out sont toujours la méthode standard pour la sélection de modèles dans le contexte des séries temporelles, en particulier pour les modèles d'apprentissage profond.

Si les données sont tirées d'un processus indexé dans l'ordre temporel, l'ensemble hors échantillon peut ne plus être indépendant de l'ensemble d'apprentissage, et l'i.i.d. la théorie du hold-out ne tient plus. Les sections suivantes visent donc à fournir des bornes de généralisation et des inégalités d'oracle pour le modèle sélectionné par minimisation de l'erreur hors échantillon dans un cadre markovien.

2.2 Modèle

2.2.1 Propriétés des chaines de Markov

Rappelons quelques définitions et propriétés des chaînes de Markov. Ils seront utiles tout au long du chapitre.

Soit $(Y_t)_{t\in\mathbb{Z}}$ une chaîne de Markov $\mathscr{Y}$, où $\mathscr{Y}$ est un espace d'état polonais.

Notons K son noyau de transition. $K(x,.)$ est la distribution de X_{n+1} conditionnée par $X_n = x$. On appelle *distribution stationnaire* la distribution de probabilité Q telle que

$$\int_{x\in\mathscr{X}} Q(dx)K(x,dz) = Q(dz).$$

Définition 2.1. *La distance variationnelle totale $d_{TV}(P,Q)$ de deux distributions P et Q définies sur le même espace d'état $(\mathscr{X},\mathscr{A})$ est définie comme*

$$d_{TV}(P,Q) := \sup_{A\in\mathscr{A}} |P(A) - Q(A)|. \tag{2.1}$$

Proposition 2.1. *Soit $(X_t)_{t\in\mathbb{Z}}$ une série à valeur dans $\mathscr{Y}$ uniformément ergodique. Il existe des constantes $C > 0$ et $0 < \rho < 1$ telles que :*

$$\sup_{x\in\mathscr{X}} d_{TV}(K^n(x,.),Q) \leq C\rho^n. \tag{2.2}$$

2.2.2 Observations

Soit $(Y_t)_{t\in\mathbb{Z}}$ une chaîne de Markov , homogène, stationnaire et uniformément ergodique, à valeurs dans $\mathscr{Y}$ espace d'état polonais. Soit Q sa distribution stationnaire. En notant p l'ordre de la chaîne de Markov, qui peut être supérieur ou égal à 1, on considère la markovianisation de la chaîne :

$$X_t := (Y_t,\dots,Y_{t-p})^T.$$

Soit $\mathscr{F}$ l'ensemble des fonctions mesurables de $\mathscr{Y}^p$ dans $\mathscr{Y}$. Une fonction $g \in \mathscr{F}$ est appelée *modèle de prédiction en une étape*, ou modèle de prédiction en abrégé.

Soit n,m des entiers positifs. On observe $n+m$ réalisations consécutives tirées du processus de Markov (X_t). Nous divisons ces observations en deux parties :

— $(X_1,\dots,X_n)$ est notre ensemble *d'entrainement* où nos modèles de prédiction sont entraînés.

— $(X_{n+1},\dots,X_{n+m})$ est notre ensemble de *validation* où nous sélectionnons le meilleur modèle.

Soulignons que, contrairement aux travaux antérieurs dans le cadre i.i.d. , l'ensemble *d'entraînement* et l'ensemble de *validation* ne sont pas supposés être indépendants. Pour résoudre cette difficulté, les travaux antérieurs introduisent généralement une certaine forme d'écart entre le train et l'ensemble de test. Par exemple, on pourrait citer la validation croisée de blocs [CM+91].

Cependant, dans nos contextes, on effectue une sélection de modèle hors échantillon classique afin d'évaluer ses performances dans le contexte de la chaîne de Markov. On introduira un écart $0 \leq b < m$ uniquement comme un outil technique qui ne doit pas apparaître à l'utilisateur de la procédure. Cet écart n'apparaît pas dans nos résultats finaux.

Définition 2.2. *Une* procédure d'apprentissage *est une fonction de* $\mathscr{Y}^n \to \mathscr{F}$. *Pour une procédure d'apprentissage* g *et une réalisation* $(x_1, \ldots, x_n)$ *de l'ensemble d'apprentissage, on notera par commodité* $\hat{g}_1^n := \hat{g}_k(x_1, \ldots, x_n)$.

Dit autrement, une procédure d'apprentissage est une méthode, qui a un certain jeu de donnés associe $(x_1, \ldots, x_n)$ un modèle de prédiction de F. Une procédure d'apprentissage classique est la minimisation du risque empirique, mais il existe de nombreuses autres procédures qui peuvent faire intervenir différentes formes de régularisation, des distributions de probabilité, etc...

2.2.3 Cadre du Machine Learning

Dans cette section, on étend le cas i.i.d. classique d'apprentissage automatique [BBL02] à des chaînes de Markov en se servant de ce qui a été présenté au chapitre 1.

Comme au chapitre 1, pour évaluer la qualité d'un modèle de prédiction, on introduit une fonction de perte $\mathrm{L} : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}^+$. Ici, on suppose que L est une fonction bornée. Sans perte de généralité, on peut alors toujours redimensionner la fonction L de telle sorte que $|\mathrm{L}(x, y)| \leq 1$. Lorsque $\mathscr{Y}$ est fini, un exemple d'une telle perte L est la fonction de perte par erreur de classification : $\mathrm{L}(x, y) = \mathbf{1}_{x \neq y}$.

Remarque 2.1. *Dans le cadre d'une régression, les pertes peuvent être non bornées (perte quadratique). Ce n'est généralement pas un problème en pratique, puisqu'il est généralement possible de restreindre l'ensemble* $\mathscr{Y}$ *à des valeurs réalistes pour des cas pratique. Dans le cadre d'une prédiction des ventes par exemple, on peut avoir* $\mathscr{Y} < 10^6$

On ré-introduit du risque théorique $\mathbb{L}(\hat{g}_1^n)$ comme dans la définition 1.1. Pour le risque théorique, on considérera dans ce chapitre qu'il s'agit du risque sur l'ensemble de validation.

$$\hat{\mathrm{L}}_m(\hat{g}_1^n) = \frac{1}{m} \sum_{k=n+1}^{n+m} \mathrm{L}(\hat{g}_1^n(\mathrm{X}_k)). \tag{2.3}$$

Étant donné que les données dans $(\mathrm{X}_t)_{n+1 \leq t \leq n+m}$ dépendent de $(\mathrm{X}_t)_{1 \leq t \leq n}$, l'espérance de l'erreur dépendra de l'indice m. C'est une différence importante avec le cadre i.i.d.[BBL02].

Définition 2.3. *Soit* $\hat{g}$ *une procédure d'apprentissage et* $\hat{g}_1^n$ *le modèle de prédiction associé entraîné sur un échantillon* $(\mathrm{X}_1 = x_1, \cdots, \mathrm{X}_n = x_n)$. *Pour un écart* $b \geq 0$, *on note* $\mathbb{L}_b(\hat{g}_1^n)$ *l'erreur attendue de* $\hat{g}_1^n$ *pour un vecteur* X_{n+1+b} *d'observations futures conditionnellement à la réalisation :*

$$\mathbb{L}_b(\hat{g}_1^n) = \mathbb{E}\left(\mathrm{L}(\hat{g}_1^n(\mathrm{X}_{n+b+1})) \,|\mathrm{X}_1 = x_1, \ldots, \mathrm{X}_n = x_n\right). \tag{2.4}$$

Notons qu'une application directe de l'ergodicité uniforme de la chaîne de Markov montre que, pour un b fini, on peut approximer l'espérance de l'erreur $\mathbb{L}_b$ par l'erreur théorique $\mathbb{L}$:

Lemme 2.1. *Soit* L *une perte bornée,* $0 \leq \mathrm{L} \leq 1$. *Soient* $\mathrm{C} \geq 1$ *et* $0 \leq \rho < 1$ *les constantes positives de l'équation* (2.2). *Avec les notations de définition 1.1, pour toute réalisation* $x_1, \cdots, x_n$ *de* $\mathrm{X}_1, \ldots, \mathrm{X}_n$ *:*

$$\left|\mathbb{L}_b(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n)\right| \leq \mathrm{C}\rho^b. \tag{2.5}$$

2.2.4 Procédures d'apprentissages

Ici, on donne rapidement les spécificités des inégalités oracle que l'on établit par rapport à ce qui a été présenté en section 1.2.2.

Soit N procédures d'apprentissage $(\hat{g}_1, \dots, \hat{g}_N)$. Il y a une différence majeure avec le cadre présenté en introduction en section 1.2.2 : ces procédures sont appliquées à l'ensemble d'apprentissage, qui est dans une relation de dépendance avec l'ensemble ou l'on sélectionne le modèle. Dit autrement, ces procédures d'apprentissage donnent N différents modèles de prédiction $\left(\hat{g}_k(x_1, \dots, x_n)\right)_{k=1,\dots,N}$ qui dépendent de l'ensemble d'apprentissage. On note également $(\hat{g}_1^n)_k := \hat{g}_k{}_1^n$ le modèle de prédiction associé à la procédure d'apprentissage $\hat{g}_k$.

Remarque 2.2. *Là encore, aucune hypothèse sur la procédure d'entraînement n'est nécessaire. En particulier, les modèles de prédiction n'ont pas besoin d'être des minimisateurs du risques empiriques, ce qui était souvent le cas pour les résultats présentés au chapitre 1.*

2.2.5 Inégalités exponentielles

Pour étudier le lien entre l'erreur empirique (2.3) et l'erreur théorique (1.4), nous avons besoin d'inégalités uniformes entre la moyenne empirique et la moyenne attendue (comme dans Lugosi [Lug02b]). Cette section vise à donner des exemples de telles inégalités que nous utiliserons par la suite. Ils sont valables pour un point de départ non aléatoire et sont dès lors bien adaptés au cadre présenté.

Tout d'abord, définissons formellement la notion de temps de mélange introduite en section 1.2.3,

Définition 2.4. *Soit $(X_t)_{t\in\mathbb{Z}}$ une chaîne de Markov homogène en temps, uniformément ergodique. Soit la distance de variation totale définie par l'équation (2.1). Le temps de mélange t_{mix} est défini par :*

$$d(t) := \sup_{x\in\mathscr{X}} \|K^t(x,\cdot) - Q\|_{TV},\ t_{mix}(\varepsilon) = \min\{t : d(t) \le \varepsilon\},\ \text{et } t_{mix} = t_{mix}(\frac{1}{4}).$$

Une adaptation directe du corollaire 2.10 et de l'équation (3.27) de Paulin [Pau15] donne la proposition suivante :

Proposition 2.2. *Soient $C \ge 1$ et $0 \le \rho < 1$ les constantes positives de l'équation (2.2). Pour toute réalisation $x_1, \cdots, x_n$ de $X_1, \cdots, X_n$ et tout écart $0 \le b < m$ entre l'ensemble d'apprentissage et de validation, avec les notations des définitions 1.1, et 2.4 :*

$$\mathbb{P}\left(\pm\left(\frac{1}{m-b}\sum_{k=n+1+b}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n)\right) > \varepsilon\right) \le \exp\left(-2\frac{(m-b)\varepsilon^2}{9t_{mix}}\right) + C\rho^b. \tag{2.6}$$

De plus, grace à l'inéquation (3.30) de Paulin [Pau15], on sait que $C\rho^b \le 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right)$, et donc que

$$\mathbb{P}\left(\pm\left(\frac{1}{m-b}\sum_{k=n+b}^{n+m-1} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n)\right) > \varepsilon\right) \le \exp\left(-2\frac{(m-b)\varepsilon^2}{9t_{mix}}\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right). \tag{2.7}$$

.

On introduit brièvement la notion de pseudo-gap spectral ; voir Paulin [Pau15] pour une présentation détaillée.

Définition 2.5. *Pour une chaîne de Markov avec noyau de transition $K(x,dz)$ et distribution stationnaire Q, nous définissons le spectre de la chaîne comme*

$$S_2 := \left\{\lambda \in \mathbb{C}\backslash 0 : (\lambda \mathbf{I} - K)^{-1} \text{ n'existe pas en tant qu'opérateur linéaire borné de } L^2(Q)\right\}.$$

Définissons également le retournement temporel de K *comme le noyau de Markov*

$$\mathrm{K}^*(x,dz) := \frac{\mathrm{K}(z,dx)}{\mathrm{Q}(dx)}\mathrm{Q}(dz).$$

Alors, l'opérateur linéaire K^* *est l'adjoint de l'opérateur linéaire* K *sur* $\mathrm{L}^2(\mathrm{Q})$. *Nous définissons une nouvelle quantité, appelée le pseudo écart spectral de* K, *comme*

$$\gamma_{ps} := \max_{k\geq 1}\left\{\gamma\left((\mathrm{K}^*)^k\mathrm{K}^k\right)/k\right\}, \tag{2.8}$$

où $\gamma\left((\mathrm{K}^*)^k\mathrm{K}^k\right)$ *désigne l'espace spectral de l'opérateur auto-adjoint* $(\mathrm{K}^*)^k\mathrm{K}^k$.

Une adaptation directe du théorème 3.4 de Paulin [Pau15] donne :

Proposition 2.3. *Soit* $(\mathrm{X}_t)_{t\in\mathbb{N}}$ *une chaîne de Markov stationnaire avec un écart spectral* γ_{ps}. *Soit* $f\in\mathrm{L}^2(\mathrm{Q})$ *avec, pour chaque* x, $|f(x)-\mathrm{E}_\mathrm{Q}(f)|\leq \mathrm{B}$. *Soit* $\mathrm{V}_f = \mathrm{Var}_\mathrm{Q}(f)$ *et* $\mathrm{S}=\sum_{i=1}^n f(\mathrm{X}_i)$. *En utilisant ces notations on a alors :*

$$\mathbb{P}\left(\pm\left(\mathrm{S}-\mathrm{E}_\mathrm{Q}(\mathrm{S})\right)>n\varepsilon\right)\leq\exp\left(-\frac{n^2\varepsilon^2\gamma_{ps}}{8(n+1/\gamma_{ps})\mathrm{V}_f+20n\varepsilon\mathrm{B}}\right). \tag{2.9}$$

De cette proposition, on déduit un lemme qui sera utilisé dans la dernière section :

Lemme 2.2. *Avec les mêmes hypothèses que la proposition précédente 2.3, pour tout* $0<\delta<1$ *:*

$$\begin{aligned}
&\mathrm{P}\left(\pm\left(\mathrm{E}_\mathrm{Q}(\mathrm{S})-\mathrm{S}\right)\leq\sqrt{\frac{8\left(\gamma_{ps}+1\right)}{\gamma_{ps}^2}n\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{20}{\gamma_{ps}}\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)\geq\\
&\mathrm{P}\left(\pm\left(\mathrm{E}_\mathrm{Q}(\mathrm{S})-\mathrm{S}\right)\leq\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{20}{\gamma_{ps}}\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)\geq 1-\delta
\end{aligned} \tag{2.10}$$

Proof Prouvons le lemme pour le signe $+$, la preuve pour le signe $-$ est la même. Par la proposition 2.3, on a pour tout $0<\delta<1$:

$$\begin{aligned}
&\mathrm{P}\left(\mathrm{E}_\mathrm{Q}(\mathrm{S})-\mathrm{S}>\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{20}{\gamma_{ps}}\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)\leq\\
&\exp\left(-\frac{\gamma_{ps}\left(\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{20}{\gamma_{ps}}\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)^2}{8\left(n+1/\gamma_{ps}\right)\mathrm{V}_f+20\mathrm{B}\left(\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{20}{\gamma_{ps}}\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)}\right)\leq\\
&\exp\left(-\frac{8(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)+40\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}\mathrm{B}\log\left(\frac{1}{\delta}\right)+\frac{1}{\gamma_{ps}}\left(20\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)^2}{8\left(n+1/\gamma_{ps}\right)\mathrm{V}_f+20\mathrm{B}\left(\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{20}{\gamma_{ps}}\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)}\right)\leq\\
&\exp\left(-\log\left(\frac{1}{\delta}\right)\frac{8(n+1/\gamma_{ps})\mathrm{V}_f+40\mathrm{B}\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{1}{\gamma_{ps}}(20\mathrm{B})^2\log\left(\frac{1}{\delta}\right)}{8\left(n+1/\gamma_{ps}\right)\mathrm{V}_f+20\mathrm{B}\left(\sqrt{\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})\mathrm{V}_f\log\left(\frac{1}{\delta}\right)}+\frac{20}{\gamma_{ps}}\mathrm{B}\log\left(\frac{1}{\delta}\right)\right)}\right)\leq\delta.
\end{aligned} \tag{2.11}$$

En notant que $\frac{8}{\gamma_{ps}}(n+1/\gamma_{ps})=\frac{8}{\gamma_{ps}^2}(\gamma_{ps}n+1)\leq\frac{8(\gamma_{ps}+1)}{\gamma_{ps}^2}n$, on termine la preuve. ■

Enfin, la proposition 2.3 et l'équation (3.27) de Paulin [Pau15] donnent la proposition suivante :

Proposition 2.4. *Soit* $(\mathrm{X}_t)_{t\in\mathbb{Z}}$ *une chaîne de Markov stationnaire avec un pseudo gap spectral* γ_{ps}. *Pour toute fonction* $g\in\mathscr{G}$, *soit* $\mathrm{V}_g=\mathrm{Var}(\mathrm{L}(g(\mathrm{X}_t))$ *la variance de la fonction de perte calculée avec*

la loi stationnaire. Pour toute réalisation $x_1,\dots,x_n$ *de* $X_1,\dots,X_n$*, et tout entier* $0 \le b < m$*, avec les notations de la définition 1.1, puisque* $C\rho^b \le 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right)$*, on vérifie :*

$$\mathbb{P}\left(\pm\left(\frac{1}{m-b}\sum_{k=n+b+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n)\right) > \varepsilon\right) \le$$
$$\exp\left(-\frac{(m-b)^2\varepsilon^2\gamma_{ps}}{8((m-b)+\frac{1}{\gamma_{ps}})V_{\hat{g}_1^n} + 20(m-b)\varepsilon}\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right). \tag{2.12}$$

2.3 Une inégalité de type Hoeffding

Dans cette section, on n'utilisera aucune autre hypothèse qu'une hypothèse de borne sur les fonctions de perte. Nous obtenons des bornes valables pour tous les modèles, mais elles peuvent être lâches dans des conditions de bruit particulières.

2.3.1 Borne exponentielle

Considérons des données hors échantillon de longueur m dans le futur de la dernière observation d'apprentissage X_n. On écrirera la borne de généralisation en prenant en compte les dernières données de validation $m-b$. Ce faisant, on omet de prendre en compte les b premières observations, mais ces observations comptent pour au plus $\frac{b}{m}$ dans l'erreur de validation empirique, et on obtient :

Proposition 2.5. *Avec les notations de la proposition 2.2, pour toute réalisation* $x_1,\dots,x_n$ *de* $X_1,\dots,X_n$*, entiers* b *et* m*, avec* $0 \le b < m$ *:*

$$\mathbb{P}\left(\pm\left(\hat{L}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n)\right) > \varepsilon + \frac{b}{m}\right) \le \exp\left(-2\frac{(m-b)\varepsilon^2}{9t_{mix}}\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right). \tag{2.13}$$

Démonstration. On prouve (2.13) pour le cas positif. Dans ce cas, $\hat{L}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) \ge \frac{b}{m}$:

$$0 \le \frac{1}{m}\sum_{k=n+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n) - \frac{b}{m} \le \frac{1}{m}\sum_{k=n+b+1}^{n+m} L(\hat{g}_1^n(X_k)) - \frac{(m-b)}{m} \times \mathbb{L}(\hat{g}_1^n) =$$
$$\frac{m-b}{m}\left(\frac{1}{m-b}\sum_{k=n+b+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n)\right) \le \frac{1}{m-b}\sum_{k=n+b+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n).$$

En utilisant la proposition 2.2 :

$$\mathbb{P}\left(\frac{1}{m}\sum_{k=n+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n) > \varepsilon + \frac{b}{m}\right) \le$$
$$\mathbb{P}\left(\frac{1}{m-b}\sum_{k=n+b+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n) > \varepsilon\right) \le \exp\left(-2\frac{(m-b)\varepsilon^2}{9t_{mix}}\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right).$$

Si $\hat{L}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) < \frac{b}{m}$, on a alors

$$\mathbb{P}\left(\frac{1}{m}\sum_{k=n+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n) > \varepsilon + \frac{b}{m}\right) = 0.$$

Ce qui fait que pour tout $g \in \mathscr{G}$,

$$\mathbb{P}\left(\hat{L}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) > \varepsilon + \frac{b}{m}\right) \le \exp\left(-2\frac{(m-b)\varepsilon^2}{9t_{mix}}\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right).$$

Cela montre l'équation (2.13) avec le signe +.

Pour le signe − , on remarque que si $\mathbb{L}(\hat{g}_1^n) - \hat{\mathrm{L}}_m(\hat{g}_1^n) \geq \frac{b}{m}$, alors :

$$0 \leq \mathbb{L}(\hat{g}_1^n) - \frac{1}{m}\sum_{k=n+1}^{n+m} \mathrm{L}(\hat{g}_1^n(\mathrm{X}_k)) - \frac{b}{m} \leq \frac{m-b}{m}\mathbb{L}(\hat{g}_1^n) - \frac{1}{m}\sum_{k=n+b+1}^{n+m} \mathrm{L}(\hat{g}_1^n(\mathrm{X}_k)) =$$
$$\frac{m-b}{m}\left(\mathbb{L}(\hat{g}_1^n) - \frac{1}{m-b}\sum_{k=n+b+1}^{n+m} \mathrm{L}(\hat{g}_1^n(\mathrm{X}_k))\right) \leq \mathbb{L}(\hat{g}_1^n) - \frac{1}{m-b}\sum_{k=n+b+1}^{n+m} \mathrm{L}(\hat{g}_1^n(\mathrm{X}_k)),$$

et, en utilisant le même raisonnement, on obtient l'équation (2.13) pour le signe −. □

En utilisant cette proposition, on peut énoncer une borne exponentielle pour la fonction de perte théorique. La preuve est donnée en annexe.

Théorème 2.1. *En utilisant les notations de la définition 2.4, pour toute réalisation $x_1, \cdots, x_n$ de $\mathrm{X}_1, \ldots, \mathrm{X}_n$, et si $1 \geq \varepsilon > 0$:*

$$\mathrm{P}\left(\pm(\hat{\mathrm{L}}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n)) > \varepsilon\right) \leq \left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + 1\right)\exp\left(-\frac{m\varepsilon^2\ln(2)}{(1+9\ln(2))\,t_{mix}}\right). \tag{2.14}$$

2.3.2 Une première inégalité oracle pour le Hold-Out

On peut maintenant établir une inégalité d'oracle, en utilisant la notation de la sous-section 1.7.

Pour établir l'inégalité d'oracle, on commencera par les inégalités entre les pertes empiriques et théoriques pour $\hat{k}$ et $\tilde{k}$. Le théorème 2.1 et la borne d'union donnent le théorème suivant :

Théorème 2.2. *Pour toute réalisation $x_1, \ldots, x_n$ de $\mathrm{X}_1, \ldots, \mathrm{X}_n$, et $1 \geq \varepsilon > 0$:*

$$\mathrm{P}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}}) > \varepsilon\right) \leq \mathrm{N}\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + 1\right)\exp\left(-\frac{m\varepsilon^2\ln(2)}{(1+9\ln(2))\,t_{mix}}\right), \tag{2.15}$$

et

$$\mathrm{P}\left(\hat{\mathrm{L}}_m((\hat{g}_1^n)_{\tilde{k}})) - \mathbb{L}((\hat{g}_1^n)_{\tilde{k}})) > \varepsilon\right) \leq \mathrm{N}\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + 1\right)\exp\left(-\frac{m\varepsilon^2\ln(2)}{(1+9\ln(2))\,t_{mix}}\right). \tag{2.16}$$

Ensuite, le théorème suivant donne une borne supérieure des espérances entre les erreurs empiriques et théoriques.

Théorème 2.3. *Notons $\beta = \frac{\ln(2)}{(1+9\ln(2))\,t_{mix}}$, et $\alpha = \left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + 1\right)$. Pour toute réalisation $(x_1, \ldots, x_n)$ de $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$, nous avons les bornes suivantes pour l'erreur de généralisation :*

$$\mathbb{E}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}})\right) \leq \sqrt{\frac{(\ln(\mathrm{N}\alpha)+1)}{\beta m}}, \tag{2.17}$$

et

$$\mathbb{E}\left(\hat{\mathrm{L}}_m((\hat{g}_1^n)_{\tilde{k}})) - \mathbb{L}((\hat{g}_1^n)_{\tilde{k}}))\right) \leq \sqrt{\frac{(\ln(\mathrm{N}\alpha)+1)}{\beta m}}. \tag{2.18}$$

Démonstration. On peut écrire :

$$\mathbb{E}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}})\right) \leq \mathbb{E}\max\left(\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}})\right), 0\right) \leq$$
$$\int_0^1 \mathrm{P}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}}) > t\right) dt \leq$$
$$\sqrt{\int_0^1 \mathrm{P}\left(\left(\max\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}}), 0\right)\right)^2 > t\right) dt} \leq \sqrt{\frac{(\ln(\mathrm{N}\alpha)+1)}{\beta m}}.$$

La preuve est sensiblement la même pour la seconde inégalité. □

Remarquons que

$$\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \mathbb{L}((\hat{g}_1^n)_{\tilde{k}}) = \mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}}) + \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\tilde{k}})) + \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\tilde{k}})) - \mathbb{L}((\hat{g}_1^n)_{\tilde{k}})), \tag{2.19}$$

et que, par définition, $\hat{\mathrm{L}}_m((\hat{g}_1^n)_{\hat{k}}) - \hat{\mathrm{L}}_m((\hat{g}_1^n)_{\tilde{k}})) \leq 0$. On a donc

$$\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \mathbb{L}((\hat{g}_1^n)_{\tilde{k}})) \leq \sup_{k\in\{1,\cdots,\mathrm{N}\}} (\mathbb{L}\left((\hat{g}_1^n)_k\right) - \hat{\mathrm{L}}_m\left((\hat{g}_1^n)_k\right)) + \sup_{k\in\{1,\cdots,\mathrm{N}\}} (\hat{\mathrm{L}}_m\left((\hat{g}_1^n)_k\right) - \mathbb{L}\left((\hat{g}_1^n)_k\right)), \tag{2.20}$$

ce qui donne l'inégalité oracle suivante :

Théorème 2.4. *Avec les notations et suppositions du Théorème 2.3, pour toute réalisation* $(x_1,\ldots,x_n)$ *de* $(\mathrm{X}_1,\ldots,\mathrm{X}_n)$*, on vérifie :*

$$\mathbb{E}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \mathbb{L}((\hat{g}_1^n)_{\tilde{k}}))\right) \leq 2\sqrt{\frac{(\ln(\mathrm{N}\alpha)+1)}{\beta m}}. \tag{2.21}$$

On peut reformuler en notant g^* *le meilleur prédicteur :*

$$\mathbb{E}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}}) - \mathbb{L}(g^*)\right) \leq \mathbb{L}((\hat{g}_1^n)_{\tilde{k}})) - \mathbb{L}(g^*) + 2\sqrt{\frac{(\ln(\mathrm{N}\alpha)+1)}{\beta m}}. \tag{2.22}$$

Notons que cette borne est du même ordre que dans le cas indépendant où m tend vers l'infini. Toutes les bornes de cette section dépendent de la constante inconnue t_{mix}, cependant, cette constante peut être estimée à partir des données (voir Wolfer et Kontorovich [WK19]).

2.4 Taux rapides avec inégalité de type Bernstein

Nous pouvons supprimer la racine carrée dans la borne du théorème 2.4 en augmentant l'estimation empirique de l'erreur de rétention d'un petit facteur constant et en utilisant une inégalité de type Berstein (comme dans Bartlett et al. [BBL02]). Lorsque l'erreur théorique est faible, les inégalités obtenues peuvent être meilleures que les inégalités précédentes.

2.4.1 Borne exponentielle pour $\mathbb{L}(\hat{g}_1^n)$

On établit des bornes exponentielles pour les erreurs empiriques légèrement modifiées et les pertes théoriques. Une application de la proposition 2.4 donne la proposition suivante, prouvée en annexe :

Proposition 2.6. *Avec les notations et hypothèses de la proposition 2.4, pour* $0<a<1$*, notons* $\mathrm{D}_a^+ = \frac{a(1+a)}{8\left(1+\frac{1}{\gamma_{ps}}\right)+20}$*, et* $\mathrm{D}_a^- = \frac{a(1-a)}{8\left(1+\frac{1}{\gamma_{ps}}\right)+20}$*. Pour toute réalisation* $(x_1,\cdots,x_n)$ *de* $(\mathrm{X}_1,\cdots,\mathrm{X}_n)$*, et toutes paires d'entiers* b *et* m*, avec* $0\leq b<m$*, on a :*

$$\mathrm{P}\left(\frac{1}{1+a}\hat{\mathrm{L}}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) > \varepsilon + \frac{b}{m}\right) \leq \exp\left(-(m-b)\gamma_{ps}\mathrm{D}_a^+\varepsilon\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right), \tag{2.23}$$

et

$$\mathrm{P}\left(\mathbb{L}(\hat{g}_1^n) - \frac{1}{1-a}\hat{\mathrm{L}}_m(\hat{g}_1^n) > \varepsilon + \frac{b}{m}\right) \leq \exp\left(-(m-b)\gamma_{ps}\mathrm{D}_a^-\varepsilon\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right). \tag{2.24}$$

En utilisant cette proposition, on peut obtenir des bornes exponentielles pour $\mathrm{P}\left(\frac{1}{1+a}\hat{\mathrm{L}}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) > \varepsilon\right)$ et $\mathrm{P}\left(\mathbb{L}(\hat{g}_1^n) - \frac{1}{1-a}\hat{\mathrm{L}}_m(\hat{g}_1^n) > \varepsilon\right)$. Ces bornes sont énoncées dans le théorème suivant. La preuve se trouve en annexe.

Théorème 2.5. *Avec les notations et hypothèses de la proposition 2.6, pour toute réalisation $x_1,\dots,x_n$ de $X_1,\dots,X_n$, $0<a<1$, et $1\geq\varepsilon>0$:*

$$P\left(\frac{1}{1+a}\hat{L}_m(\hat{g}_1^n)-\mathbb{L}(\hat{g}_1^n)>\varepsilon\right)\leq\left(1+2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\right)\exp\left(-\frac{D_a^+m\varepsilon}{4t_{mix}}\right), \tag{2.25}$$

et

$$P\left(\mathbb{L}(\hat{g}_1^n)-\frac{1}{1-a}\hat{L}_m(\hat{g}_1^n)>\varepsilon\right)\leq\left(1+2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\right)\exp\left(-\frac{D_a^-m\varepsilon}{4t_{mix}}\right). \tag{2.26}$$

2.4.2 Borne sur l'erreur de généralisation

Avec les notations de la section 2.2.4, on considère $(\hat{g}_1^n)_{k=1,\cdots,N}$, une collection finie de classifieurs obtenus en traitant une réalisation d'échantillon d'apprentissage de longueur n. Le théorème 2.5 et la borne d'union donnent le corollaire suivant :

Corollaire 2.1. *Pour toute réalisation $x_1,\dots,x_n$ de $X_1,\dots,X_n$, $0<a<1$ et $0<\varepsilon\leq 1$:*

$$P\left(\frac{1}{1+a}\hat{L}_m((\hat{g}_1^n)_{\tilde{k}}))-\mathbb{L}((\hat{g}_1^n)_{\tilde{k}}))>\varepsilon\right)\leq N\left(1+2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\right)\exp\left(-\frac{D_a^+m\varepsilon}{4t_{mix}}\right), \tag{2.27}$$

et

$$P\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})>\varepsilon\right)\leq N\left(1+2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\right)\exp\left(-\frac{D_a^-m\varepsilon}{4t_{mix}}\right), \tag{2.28}$$

Puis, dans le théorème suivant, on donne une borne supérieure des espérances de ces expressions :

Théorème 2.6. *Notons $\alpha=\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right)+1\right)$, pour toute réalisation $(x_1,\cdots,x_n)$ de $(X_1,\cdots,X_n)$, on a les bornes suivantes pour l'erreur de généralisation :*

$$\mathbb{E}\left(\frac{1}{1+a}\hat{L}_m((\hat{g}_1^n)_{\tilde{k}}))-\mathbb{L}((\hat{g}_1^n)_{\tilde{k}}))\right)\leq\frac{4t_{mix}(\ln(N\alpha)+1)}{D_a^+m}, \tag{2.29}$$

et

$$\mathbb{E}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})\right)\leq\frac{4t_{mix}(\ln(N\alpha)+1)}{D_a^-m}, \tag{2.30}$$

Démonstration. On peut écrire :

$$\mathbb{E}\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})\right)\leq\mathbb{E}\max\left(\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})\right),0\right)=$$
$$\int_0^1 P\left(\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})>t\right)dt\leq\frac{4t_{mix}(\ln(N\alpha)+1)}{D_a^-m}.$$

La preuve de la seconde inégalité est symétrique. □

Remarquons

$$\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\mathbb{L}((\hat{g}_1^n)_{\tilde{k}}))=$$
$$\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})+\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})$$
$$-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\tilde{k}}))+\frac{1}{1+a}\hat{L}_m((\hat{g}_1^n)_{\tilde{k}}))-\mathbb{L}((\hat{g}_1^n)_{\tilde{k}})+2\frac{a}{1-a^2}\hat{L}_m((\hat{g}_1^n)_{\tilde{k}},$$

avec, par définition, $\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\hat{k}})-\frac{1}{1-a}\hat{L}_m((\hat{g}_1^n)_{\tilde{k}}))\leq 0$. Du coup, on a

$$\mathbb{L}((\hat{g}_1^n)_{\hat{k}})-\mathbb{L}((\hat{g}_1^n)_{\tilde{k}}))\leq$$
$$\sup_{k\in\{1,\cdots,N\}}\left(\mathbb{L}\left((\hat{g}_1^n)_k\right)-\frac{1}{1-a}\hat{L}_m\left((\hat{g}_1^n)_k\right)\right)+$$
$$\sup_{k\in\{1,\cdots,N\}}\left(\frac{1}{1+a}\hat{L}_m\left((\hat{g}_1^n)_k\right)-\mathbb{L}\left((\hat{g}_1^n)_k\right)\right)+2\frac{a}{1-a^2}\hat{L}_m((\hat{g}_1^n)_{\tilde{k}},$$

et on obtient les inégalités suivantes

Théorème 2.7. *Avec les notations du Théorème 2.6, pour toute réalisation* $(x_1,\dots,x_n)$ *de* $(\mathrm{X}_1,\dots,\mathrm{X}_n)$*, on a :*

$$\mathbb{E}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)\right)\leq \frac{4t_{mix}\left(\ln\left(\mathrm{N}\alpha\right)+1\right)}{\mathrm{D}_a^- m}+\frac{4t_{mix}\left(\ln\left(\mathrm{N}\alpha\right)+1\right)}{\mathrm{D}_a^+ m}+\frac{2a}{1-a^2}\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right). \tag{2.31}$$

Ou, si l'on note g^* *le meilleur prédicteur théorique :*

$$\mathbb{E}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)\right)\leq\left(1+\frac{2a}{1-a^2}\right)\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)+$$
$$\frac{4t_{mix}\left(\ln\left(\mathrm{N}\alpha\right)+1\right)}{\mathrm{D}_a^- m}+\frac{4t_{mix}\left(\ln\left(\mathrm{N}\alpha\right)+1\right)}{\mathrm{D}_a^+ m}+\frac{2a}{1-a^2}\mathbb{L}(g^*).$$

Cet inégalité diffère de l'inégalité d'oracle (1.7) ; ces bornes sont meilleures que les bornes de la section précédente si l'erreur théorique $\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)$ est nulle et m est assez grand. Toutes les bornes de cette section dépendent des constantes inconnues t_{mix} et γ_{ps}, mais elles peuvent être estimées à partir des données (voir Wolfer et Kontorovich [WK19]).

2.5 Vitesse rapide sous condition de bruit

La condition précédente $\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)=0$ peut être considérée comme une condition de bruit grossier. On peut essayer d'affiner l'analyse car le hold-out bénéficie d'excellentes propriétés théoriques lorsque l'on ajoute des conditions de bruit dans le cas i.i.d.(voir Blanchard et Massart [BM06], Boucheron et al. [BOL05] ou Massart [Mas03]). Par conséquent, on étudiera ses propriétés dans le cas de Markov dans des conditions similaires pour le bruit. Tout d'abord, présentons l'hypothèse sur le bruit.

Hypothèse de bruit (H) :

— Il existe une fonction $\omega(\cdot)$ telle que $\omega(x)/\sqrt{x}$ soit non-croissante et, pour toute fonction $g\in\mathscr{G}$,

$$\sqrt{\mathrm{Var}\left(\mathbf{1}_{g\neq g^*}\right)}\leq\omega(\mathbb{L}(g)-\mathbb{L}(g^*)),$$

les espérances étant calculé selon la loi stationnaire de la chaine de Markov $(\mathrm{X}_t)_{t\in\mathbb{Z}}$.

— Soit τ_m^* la plus petite solution positive de $\omega(\varepsilon)=\sqrt{m}\varepsilon$.

Avec les notations de la section 2.2.4, on considère $\left(\left(\hat{g}_1^n\right)_k\right)_{k=1,\cdots,\mathrm{N}}$, une collection finie de fonctions de prédiction obtenu en traitant une jeu d'échantillons d'apprentissage de longueur n. On peut alors poser la proposition suivante. Elle sera démontré en annexe

Proposition 2.7. *Si l'hypothèse de bruit* (H) *est vérifiée, alors pour toute réalisation* $x_1,\dots,x_n$ *de* $\mathrm{X}_1,\dots,\mathrm{X}_n$*, tout couple d'entiers* b *et* m *tels que* $0\leq b<m$ *et tout* $\theta\in]0,1[$ *:*

$$\mathbb{P}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)-(1+\theta)\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)>\varepsilon+\frac{(1+\theta)2b}{m}\right)\leq$$
$$\mathrm{N}\exp\left(-\frac{1}{1+\theta}\frac{\theta\gamma_{ps}(m-b)}{16(1+\frac{1}{\gamma_{ps}})m\tau_m^*+80\theta}\varepsilon\right)+2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right). \tag{2.32}$$

2.5.1 Une borne exponentielle sous condition de bruit

On peut maintenant énoncer une borne exponentielle sous condition de bruit. Le théorème suivant est démontré en annexe.

Théorème 2.8. *Si l'hypothèse de bruit* (**H**) *est vérifiée, alors pour toute réalisation* $x_1,\cdots,x_n$ *de* $X_1,\cdots,X_n$, *et tout* $\theta\in]0,1[$ *:*

$$\mathbb{P}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)-(1+\theta)\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)>\varepsilon\right)\leq$$
$$\left(N+2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\right)\exp\left(-\frac{1}{4t_{mix}(1+\theta)}\frac{\theta\gamma_{ps}m}{16(1+\frac{1}{\gamma_{ps}})m\tau_m^*+80\theta}\varepsilon\right). \tag{2.33}$$

En intégrant cette expression, on peut obtenir finalement l'inégalité oracle que l'on recherchait.

Corollaire 2.2. *Si l'hypothèse sur le bruit* (**H**) *est vérifiée,on obtient alors pour toute réalisation* $x_1,\cdots,x_n$ *de* $X_1,\cdots,X_n$, *et tout* $\theta\in]0,1[$ *:*

$$\mathbb{E}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)\right)\leq(1+\theta)\times$$
$$\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)+\frac{4t_{mix}(16(1+\frac{1}{\gamma_{ps}})m\tau_m^*+80\theta)}{\theta\gamma_{ps}m}\left(\ln\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right)+N\right)+1\right)\right) \tag{2.34}$$

Remarque 2.3. HYPOTHÈSE DE MAMMEN-TSYBAKOV : *Supposons que la condition de bruit de Mammen-Tsybakov est vérifiée pour un certain exposant* α *, c'est-à-dire que l'on peut choisir* $w(r)=\left(\frac{r}{h}\right)^{\alpha/2}$ *pour une certaine valeur h positive. Alors,* $\tau_m^*=\left(mh^\alpha\right)^{-1/(2-\alpha)}$*, et le corollaire précédent devient :*

$$\mathbb{E}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)\right)\leq(1+\theta)\times\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)+$$
$$\frac{4t_{mix}(16(1+\frac{1}{\gamma_{ps}})h^{-\alpha/(2-\alpha)}m^{1-1/(2-\alpha)}+80\theta)}{\theta\gamma_{ps}m}\left(\ln\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right)+N\right)+1\right)= \tag{2.35}$$
$$(1+\theta)\times\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)+$$
$$\left(\frac{320t_{mix}}{\gamma_{ps}m}+\frac{4t_{mix}\left(16(1+\frac{1}{\gamma_{ps}})h^{-\alpha/(2-\alpha)}\right)}{\theta\gamma_{ps}m^{1/(2-\alpha)}}\right)\left(\ln\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right)+N\right)+1\right) \tag{2.36}$$

Remarque 2.4. *Si l'espace d'état* $\mathcal{Y}$ *est égal à* $\{0,1\}$*, et si la fonction de prédiction* $\eta(Y_{t-1},\ldots,Y_{tp})=\mathbb{E}\left(Y_t\middle|Y_{t-1},\cdots,Y_{tp}\right)$ *est telle que pour tout* $y_{t-1},\ldots,y_{tp}\in\mathcal{Y}^p$*,* $\left|2\eta(y_{t-1},\cdots,y_{tp})-1\right|>h$*, alors la condition de bruit de Mammen-Tsybakov est vérifiée avec* $\alpha=1$*.*

$$\mathbb{E}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)\right)\leq(1+\theta)\times\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)+$$
$$\left(\frac{320t_{mix}}{\gamma_{ps}m}+\frac{4t_{mix}\left(16(1+\frac{1}{\gamma_{ps}})h^{-\alpha/(2-\alpha)}\right)}{\theta\gamma_{ps}m}\right)\left(\ln\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right)+N\right)+1\right).$$

Donc, si $\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)=0$*, alors on obtient une vitesse rapide pour la convergence de l'inégalité oracle*

$$\mathbb{E}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)\right)\leq O\left(\frac{1}{m}\right).$$

Remarque 2.5. *Soit* $p\geq 1$ *un entier fixe. Considérons l'ensemble des chaînes de Markov ergodiques homogènes d'ordre p avec* $\mathcal{Y}=\{0,1\}$*. Posons la mesure uniforme sur les composantes des matrices de transition possibles de cet ensemble. La mesure de Lebesgue de l'ensemble des modèles pour lesquels la condition de bruit de Mammen-Tsybakov ne tient pas avec* $\alpha=1$ *sera nulle. Par conséquent, pour presque tous les modèles, nous obtenons le taux précédent pour l'inégalité d'oracle.*

Chapitre 3

Des inégalités de concentration pour des champs de vecteurs non-causaux

Ce chapitre est issu d'un article écrit en collaboration avec Raphaël Langhendries accepté avec révisions dans Electronic Journal of Statistics, sous le titre "Concentration inequalities for non-causal random fields". Il propose un cadre de modélisation des champs aléatoires non causals et démontre une inégalité de concentration de type Hoeffding dans ce cadre. Au passage, il établit également une inégalité oracle pour ce cadre. La preuve de ce résultat repose sur une approximation locale du champ aléatoire non causal par une fonction d'un nombre fini de i.i.d. de variables aléatoires.

3.1 Motivation

Comme on l'a présenté en section 1.2.3, le cadre des chaînes de Markov est un cadre classique pour établir des inégalités de concentration. Rappelons l'équation générale d'une chaîne de Markov (X_t)

$$X_t = F(X_{t-1}, \dots, X_{t-i}, \varepsilon_t), \text{ pour une fonction F et des innovations i.i.d. } (\varepsilon_t)_{t\in\mathbb{Z}}. \tag{3.1}$$

De nombreuses approches ont été utilisées pour établir des inégalités de concentration dans le cadre de chaine de Markov. Ces approches incluent la technique de renouvellement (voir : [BC19]), le couplage de Marton dans [Pau15] et l'approche par martingale (voir [ADF19; CW18; DDF19; DF15; KR08]).

Cependant, les chaînes de Markov ne sont pas suffisantes pour modéliser n'importe quel type de dépendance, même si on se restreint à des dépendances locales. En effet, certaines données peuvent présenter une dépendance non causale. Dans le cas unidimensionnel, cela signifie qu'un point de données ne dépend pas seulement de points de données passés, mais également de données futures.

Par exemple, cette situation se produit pour les données textuelles. Une tâche clé impliquant des données textuelles est le **problème de complétion**. Il consiste à remplir des blancs dans un texte en utilisant des mots environnants. Cette tâche est réalisée en créant un modèle de langage, c'est-à-dire une distribution de probabilité de mots pour un contexte donné (mots passés et futurs). Dans un texte, il est naturel de modéliser la distribution d'un mot étant donné le contexte en se servant non seulement des mots passés, mais également du futur . Les données textuelles sont un exemple typique de données générées par un processus **non causal**.

En pratique, des modèles non causals sont déjà utilisés pour apprendre des modèles de langage. Parmi ces modèles, il faut citer les réseaux de neurones bidirectionnels [SP97]. Ces derniers ont récemment reçu beaucoup d'attention pour leurs performances en traitement automatique du langage naturel. En particulier, le modèle BERT [KT19] est devenu un incontournable pour un très large éventail de tâches de traitement du langage naturel, telles que la traduction, le balisage de

parties de discours, l'analyse des sentiments. Cependant, malgré leur succès dans les applications pratiques, il manque un cadre théorique pour analyser de tels modèles non-causaux.

Si la dimension du réseau de variables aléatoires augmente, on obtient un **champ aléatoire**. L'extension naturelle des chaînes de Markov aux champs aléatoires conduit à des champs aléatoires causals : ici la dépendance se propage selon des directions préférentielles (voir [DPRS17] pour un exemple d'application). Les champs aléatoires non causals apparaissent naturellement dans de nombreuses applications. Par exemple, on peut modéliser de la génération d'image ou de texture par un champ aléatoire non causal défini sur un réseau à deux dimensions. Dans ce cas, le problème de complétion consiste à remplir les pixels manquants en utilisant les pixels voisins [BBC+01]. On utilise les informations de chaque direction (haut, bas, gauche, droite) pour pouvoir compléter le pixel manquant. Cependant, contrairement à ce qui se passe dans le cas des données textuelles, de telles techniques de complétion de pixels ne sont pas dominantes pour le traitement d'images. Une autre application d'importance est le cas de complétion des ensembles de données géographiques qui ont du sens pour un cadre écologique (voir `http://doukhan.u-cergy.fr/EcoDep.html`), pour lesquels les applications sont d'une importance fondamentale.

3.1.1 Plan du chapitre

Dans la section 3.2, on présente les champs aléatoires non-causaux proposés dans [DT07]. Cette section présente les hypothèses qui seront utilisées dans le reste du chapitre. On donne également quelques exemples de modèles satisfaisant notre cadre, puis nous définissons la statistique $S_{\mathscr{J}}$ d'intérêt pour laquelle nous cherchons à prouver une inégalité de concentration. Les principaux résultats peuvent être trouvés dans la section 3.3 ainsi que quelques applications immédiates à l'apprentissage automatique et des comparaisons avec d'autres inégalités de concentration.

Le reste du chapitre est consacré à la preuve des principaux résultats.
Dans la section 3.4, on introduit une approximation $\tilde{S}_{\mathscr{J}}^{[d]}$ de $S_{\mathscr{J}}$ qui ne dépend que d'un nombre *fini* de variables indépendantes. On établit également un résultat évaluant la qualité de cette approximation.
Ensuite, dans la section 3.5 on prouve la principale inégalité de concentration en utilisant l'inégalité de McDiarmid proposée par [Com15] sur $\tilde{S}_{\mathscr{J}}^{[d]}$.

3.2 Modèle

Dans cette section, on présente un modèle inspiré de [DT07] afin de considérer un champ aléatoire non causal. Ensuite, on présente les hypothèses et la statistique cible $S_{\mathscr{J}}$ pour laquellle nous désirons montrer une inégalité de concentration.

3.2.1 Définitions et notations

Introduisons quelques notions qui resteront valables tout au long de ce chapitre. Désormais, toutes les variables aléatoires seront définies sur un espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$.

Dimension d'un champ de vecteurs Soit $\kappa \in \mathbb{N}$ la *dimension* du champ aléatoire d'intérêt.

Cette **dimension** ne doit pas être confondue avec la dimension classique qui apparaît dans les statistiques de grande dimension, c'est-à-dire le nombre de paramètres du modèle. Ici, le nombre de paramètres augmente de façon exponentielle avec la dimension κ.

Cadre probabiliste Soit $\mathscr{X}$ un espace de Banach doté d'une norme $\|\cdot\|$. On définit la $m-$norm $\|X\|_m$ d'une variable aléatoire X comme

$$\|X\|_m = (\mathbb{E}\|X\|^m)^{\frac{1}{m}} \quad \text{with } \|\cdot\| \text{ est une norme sur } \mathscr{X}.$$

Nous utilisons également la norme uniforme $\|X\|_\infty = \inf\{C, \|X\| \leq C \text{ p.s. }\} = \lim_{m\to\infty} \|X\|_m$.

Voisinages $\mathcal{V}(\delta, s)$ et $\mathscr{B}$. Soit δ un κ-uplet d'entier non négatif, c'est-à-dire $\delta = (\delta_1, \dots, \delta_\kappa) \in \mathbb{N}^\kappa$ et s un κ-uplet d'entiers, c'est-à-dire $s = (s_1, \dots s_\kappa) \in \mathbb{Z}^\kappa$. Le voisinage δ de s, $\mathcal{V}(\delta, s)$ est défini comme l'orthotope κ suivant

$$\mathcal{V}(\delta, s) = \{t = (t_1, \dots, t_\kappa) \in (\mathbb{Z})^\kappa / \, \forall i, s_i - \delta_i \leq t_i \leq s_i + \delta_i\}.$$

Par la suite, on étudiera un voisinage spécifique $\mathscr{B} = \mathcal{V}(\delta, 0) \backslash \{0\}$ pour une valeur fixe de δ.
De plus on notera n_B le cardinal de $\mathcal{V}(\delta, 0) = \mathscr{B} \bigcup \{0\}$ (donc $n_B = card(\mathscr{B}) + 1$).

Innovations $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$. Soit $(\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$ un champ aléatoire indexé par $\mathbb{Z}^\kappa$ de variables aléatoires *indépendantes et identiquement distribuées* ε_t sur un espace de Banach E. Pour raccourcir la notation, on note $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$ l'ensemble du champ aléatoire.
De plus, on définit μ_ε la distribution de probabilité d'une variable aléatoire ε_t et $\mu = \bigotimes_{t \in \mathbb{Z}^\kappa} \mu_\varepsilon$, le produit de ces distributions sur $E^{\mathbb{Z}^\kappa}$. μ est la distribution de $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$.

3.2.2 Relations non-causales locales

Dans cet chapitre, on étudiera le champ aléatoire non causal de dimension κ $(X_t)_{t \in \mathbb{Z}^\kappa}$. Ce champ aléatoire présente une *dépendance locale* : pour chaque $t \in \mathbb{Z}^\kappa$, X_t dépend de son voisinage indexé par $\mathcal{V}(\delta, t)$ et sur l'innovation ε_t.
Formellement, on suppose qu'il existe une fonction $F : \mathscr{X}^{\mathscr{B}} \times E \mapsto \mathscr{X}$ telle que $(X_t)_{t \in \mathbb{Z}^\kappa}$ est une solution stationnaire de l'équation suivante

$$X_t = F((X_{t+s})_{s \in \mathscr{B}}, \varepsilon_t) \qquad \forall t \in \mathbb{Z}^\kappa. \tag{3.2}$$

[DT07] assure l'existence et l'unicité de telles solutions donnant une hypothèse de contraction sur F. Ici, on suppose seulement que l'on a un champ aléatoire (fortement) stationnaire $(X_t)_{t \in \mathbb{Z}^\kappa}$ vérifiant cette équation et sans soucier de son unicité.

Par la suite, on note μ_X la distribution stationnaire d'une variable aléatoire marginale X_t. On suppose que les lois μ_X et μ_ε sont stables par F. Cela signifie que, si $(X_t)_{t \in \mathscr{B}}$ est tirée avec une distribution marginale μ_X et ε est tirée avec la distribution μ_ε, alors la variable aléatoire $F((X_s)_{s \in \mathscr{B}}, \varepsilon)$ suit la distribution μ_X.

Pour $s \in \mathscr{B}$, X_{t+s} dépend de X_t, mais X_t dépend aussi de X_{t+s}. Par conséquent, il n'est plus possible de décrire $(X_t)_{t \in \mathbb{Z}^\kappa}$ à la suite d'un processus de martingale. C'est pourquoi, nous appelons $(X_t)_{t \in \mathbb{Z}^\kappa}$ un champ aléatoire non-causal.

Remarque 3.1. *Ceci est similaire aux cadres des séries chronologiques auto-régressives dans le cas unidimensionnel présenté aux chapitres précédents. C'est pourquoi, notre cadre peut être vu comme une généralisation des chaînes de Markov stationnaires. En effet lorsque* $\delta = \kappa = 1$ *et* $\mathscr{B} = \{-1\}$, $(X_t)_{t \in \mathbb{Z}}$ *est une chaîne de Markov homogène et stationnaire .*

Hypothèse de contraction

Dans cette section, on étend les conditions de contraction présentés dans le cadre d'une chaine de Markov en section 1.2.3 à des séries temporelles définies par l'équation (3.2). On définit d'abord une condition de contraction absolue.

Définition 3.1. CONTRACTION ABSOLUE *Il existe* $(\lambda_t)_{t \in \mathscr{B}}$, *tel que* $\rho := \sum_{t \in \mathscr{B}} \lambda_t < 1$, *et pour tout tuples* $\mathscr{Y} = (y_t)_{t \in \mathscr{B}}$ *et* $\mathscr{Y}' = (y'_t)_{t \in \mathscr{B}}$ *à valeurs dans* $\mathscr{X}$ *indexés par* $\mathscr{B}$ *et pour tout* $\varepsilon \in E$.

$$\|F(\mathscr{Y}, \varepsilon) - F(\mathscr{Y}', \varepsilon)\| \leq \sum_{t \in \mathscr{B}} \lambda_t \|y_t - y'_t\|. \tag{3.3}$$

Cette condition est similaire à la condition proposée par [DT07]. Cette condition est forte, et n'est pas satisfaite par de nombreux modèles habituels.

On veut donc assouplir cette condition. Pour cela, on considére que la contraction n'est vérifiée que pour un moment d'ordre de m. Cela conduit à l'hypothèse suivante.

Définition 3.2. CONTRACTION FAIBLE (HYPOTHÈSE $\mathbf{H_1^m}$). *Soit $m \in \mathbb{N}$. Il existe $(\lambda_t)_{t\in\mathscr{B}} \in [0,1]^{\mathscr{B}}$ tel que, pour tout couplage de variables aléatoires $(Y_t)_{t\in\mathbb{Z}^\kappa}, (Y'_t)_{t\in\mathbb{Z}^\kappa}$ à valeur dans $\mathscr{X}$ de distribution marginale μ_X et tout champ de vecteurs i.i.d $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t\in\mathbb{Z}^\kappa}$ de distribution produit μ.*

1. $\forall t \in \mathbb{Z}^\kappa \|Y_t - Y'_t\|$ *admet un moment d'ordre m,*
2. $\forall t \in \mathbb{Z}^\kappa, \|F((Y_{t+s})_{s\in\mathscr{B}}, \varepsilon_t) - F((Y'_{t+s})_{s\in\mathscr{B}}, \varepsilon_t)\|_m \leq \sum_{s\in\mathscr{B}} \lambda_s \|Y_{t+s} - Y'_{t+s}\|_m$.
3. $\sum_{t\in\mathscr{B}} \lambda_t < 1$ *(On notera par la suite* $\rho = \sum_{t\in\mathscr{B}} \lambda_t$*),*

Il s'agit d'une condition de contraction "en moment" pour un unique moment m. Pour des raisons qui apparaîtront par la suite, en général, on aimerait que $m > 4$. Cette condition ne demande pas que la fonction F soit contractante, comme dans l'équation (3.3). Il s'agit d'une condition complexe à vérifier sur un nouveau modèle car l'hypothèse "pour tout couplage" permet tout sorte de dépendance entre (Y_t), (Y'_t) et ε_t. Cette condition est cependant nécessaire à cause de la nature non causale des champs de vecteurs considérés.

L'hypothèse de contraction faible conduit à des résultats plus faibles que la contraction absolue. On utilise donc aussi le compromis suivant : Lorsque m passe à ∞, ($\mathbf{H_1^m}$) devient :

Définition 3.3. CONTRACTION UNIFORME (HYPOTHÈSE $\mathbf{H_1^\infty}$). *Il existe $(\lambda_t)_{t\in\mathscr{B}} \in [0,1]^{\mathscr{B}}$ tel que, pour tout couplages de variables aléatoires $(Y_t)_{t\in\mathbb{Z}^\kappa}, (Y'_t)_{t\in\mathbb{Z}^\kappa}$ de distribution marginale μ_X and tout champ de vecteur $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t\in\mathbb{Z}^\kappa}$ de distribution produit μ.*

1. $\forall t \in \mathbb{Z}^\kappa, \|Y_t - Y'_t\|$ *admet un moment d'ordre* ∞
2. $\forall t \in \mathbb{Z}^\kappa, \|F((Y_{t+s})_{s\in\mathscr{B}}, \varepsilon_t) - F((Y'_{t+s})_{s\in\mathscr{B}}, \varepsilon_t)\|_\infty \leq \sum_{s\in\mathscr{B}} \lambda_s \|Y_{t+s} - Y'_{t+s}\|_\infty$.
3. $\sum_{t\in\mathscr{B}} \lambda_t < 1$

La contraction absolue d'un champ de vecteur entraîne à la fois la contraction uniforme et faible.

Lemme 3.1. *Si* F *vérifie la contraction absolue et si la première condition de la condition de contraction faible (respectivement uniforme), alors il vérifie la condition de contraction faible (resp. uniforme).*

La première condition de contraction faible et uniforme est immédiatement vérifiée dès que $\mathscr{X}$ est borné. Cependant, ce n'est pas une condition nécessaire, en particulier pour ($\mathbf{H_1^m}$). En particulier, sa première condition est vérifiée dès que μ_X est à queue courte.

Le lien entre contraction forte et faible est plus compliqué à établir. Si on a la contraction uniforme, alors pour tout couplage, on a la condition 2. de l'hypothèse ($\mathbf{H_1^m}$) à partir d'un certain rang m. Mais on n'est pas parvenu à établir s' il existe un tel rang m pour tout couplage.

Condition de couplage

On introduit une hypothèse de couplage similaire à une utilisation dans [DF15]. Elle contrôle le moment de la différence entre deux variables indépendantes suivant la même distribution μ_X.

Définition 3.4. COUPLAGE BORNÉ (HYPOTHÈSE $\mathbf{H_2^m}$) *Soit $m \in \mathbb{N} \cup \{\infty\}$. On suppose que pour deux variables aléatoires indépendantes* X, Y, *à valeurs dans $\mathscr{X}$ et de distribution marginale μ_X, il existe une constante $\mathbb{V}_m$ telle que :*

$$\|Y - X\|_m \leq \mathbb{V}_m.$$

Cette hypothèse est immédiatement vérifiée dès que $diam(\mathscr{X}) \leq \infty$ (car $\forall m, \mathbb{V}_m < diam(\mathscr{X})$). De plus, si $(\mathbf{H_1^m})$ est vérifié, alors $(\mathbf{H_2^m})$ l'est aussi. Néanmoins, la quantité $\mathbb{V}_m$ joue un rôle important dans les inégalités de concentration et peut être significativement plus petite que $diam(\mathscr{X})$.

Remarque 3.2. *Si $m, m' \in \mathbb{N} \cup \{\infty\}$ et $m' \leq m$, alors* $(\mathbf{H_2^m}) \Longrightarrow (\mathbf{H_2^{m'}})$.

3.2.3 Fonction d'intéret Φ et statistique $S_{\mathscr{J}}$

Dans cette section, on définira la statistique que l'on cherchera à contrôler au moyen d'une inégalité de Hoeffding.

Tout au long de ce chapitre, on s'intéressera à une fonction $\Phi : \mathscr{X}^{\bar{\mathscr{B}}} \mapsto \mathbb{R}$ définie sur un petit voisinage :

$$\bar{\mathscr{B}} = \mathscr{V}(\bar{\delta}, 0) = \prod_{i=1}^{\kappa} [-\bar{\delta}_i, \bar{\delta}_i] \tag{3.4}$$

et on note $n_{\bar{\mathrm{B}}}$ le cardinal de $\bar{\mathscr{B}}$. Il faut noter que les définitions de $\mathscr{B}$ et $\bar{\mathscr{B}}$ sont légèrement différentes, le point $\{0\}$ est inclus dans l'ensemble $\bar{\mathscr{B}}$ et non dans $\mathscr{B}$. En effet, $\mathscr{B}$ est utilisée pour représenter la dépendance, et ne peut donc pas contenir le point qu'on veut calculer, alors que $\bar{\mathscr{B}}$ est un voisinage choisi empiriquement pour l'estimation et pour calculer une perte. On doit donc pouvoir calculer une perte, ce qui donne généralement $\Phi((\mathrm{X}_{s+t})_{t\in\bar{\mathscr{B}}}) = \mathrm{L}(g((\mathrm{X}_{t+s})_{t\in\bar{\mathscr{B}}, t\neq 0}), \mathrm{X}_s)$ si on prend des notations analogue au chapitre 1.

Les définitions de n_{B} et $n_{\bar{\mathrm{B}}}$ sont comparables, en effet $n_{\mathrm{B}} = card(\mathscr{B}) + 1$ et $n_{\bar{\mathrm{B}}} = card(\bar{\mathscr{B}})$. Ensuite, pour un sous-ensemble $\mathscr{J}$ donné d'indices, on définit la statistique $S_{\mathscr{J}}$

$$S_{\mathscr{J}} = \sum_{s\in\mathscr{J}} \Phi((\mathrm{X}_{s+t})_{t\in\bar{\mathscr{B}}}). \tag{3.5}$$

On rappelle d'abord que le champ de vecteurs est stationnaire donc $\mathbb{E}[\Phi((\mathrm{X}_{s+t})_{t\in\bar{\mathscr{B}}})] = \mathbb{E}[S_{\mathscr{J}}]$ est bien défini. L'objectif est de contrôler la différence entre $S_{\mathscr{J}}$ et $\mathbb{E}[S_{\mathscr{J}}]$, ce qu'on appelle la déviation de $S_{\mathscr{J}}$. On introduit ci-dessous une hypothèse concernant cette fonction Φ.

Définition 3.5. L-LIPSCHITZ SÉPARABILITÉ (HYPOTHÈSE $\mathbf{H_3}$) *Φ est L-Lipschitz separable pour une certaine valeur $\mathrm{L} > 0$ si, pour tout couple de jeu de valeurs $(u_t)_{t\in\bar{\mathscr{B}}}$, $(v_t)_{t\in\bar{\mathscr{B}}}$ de $\mathscr{X}^{\bar{\mathscr{B}}}$:*

$$\|\Phi((u_t)_{t\in\bar{\mathscr{B}}}) - \Phi((v_t)_{t\in\bar{\mathscr{B}}})\| \leq \mathrm{L} \sum_{t\in\bar{\mathscr{B}}} \|u_t - v_t\|. \tag{3.6}$$

Cette hypothèse est analogue à celle proposée par [DF15] dans un cadre dépendant causal. De plus, des travaux récents (voir [VS18]) suggèrent qu'une telle hypothèse est appropriée pour traiter les modèles d'apprentissage en profondeur et fournit un algorithme pour estimer la constante de Lipschitz.
Pour simplifier, on supposera tout au long de ce chapitre que cette hypothèse est vérifiée pour $\mathrm{L} = 1$.

Définition 3.6. HYPOTHÈSE $\mathbf{H_4}$ *On suppose que Φ est bornée.*

$$\forall x \in \mathscr{X}^{\bar{\mathscr{B}}}, |\Phi(x)| \leq \mathrm{M}. \tag{3.7}$$

Cette hypothèse est immédiatement vérifiée lorsque $\mathscr{X}$ est borné. Le cas de $\mathscr{X}$ illimité sera discuté dans la section 3.3.2.

3.3 Résultats principaux

Dans cette section, on présentera les résultats principaux du chapitre, à savoir l'extension de l'inégalité de Hoeffding depuis le cadre i.i.d (Theorème 1.2) vers le cadre introduit non-causal en section précédente .

3.3.1 Inégalités de concentration et borne de deviation en espérance pour les champs aléatoires non causals

On présente ci-dessous des versions simplifiées des inégalités obtenus, notamment en ne détaillant pas les constantes. Les théorèmes complets peuvent être trouvés dans la section 3.5.2. Présentons d'abord le cas de la contraction uniforme.

Théorème 3.1. INÉGALITÉ TYPE-HOEFFDING, CADRE NON CAUSAL, CONTRACTION UNIFORME *Soit* $n = Card(\mathscr{I})$, $n_B = Card(\mathscr{B}) + 1$ *et* $n_{\bar{B}} = Card(\bar{\mathscr{B}})$. *Si* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés, alors il existe une constante* A *telle que, pour* $\varepsilon > 2n_{\bar{B}}\mathbb{V}_\infty$,

$$\mathbb{P}\left(|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\left(\varepsilon - 2n_{\bar{B}}\mathbb{V}_\infty\right)^2}{\left(n_{\bar{B}}\mathbb{V}_\infty\right)^2\left(1 + An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right).$$

Avec A *telle que* $\Upsilon(d)^2 d^\kappa \leq A(\kappa!)^2\lceil\ln(n)\rceil^\kappa$, *où* Υ *est la fonction définie au Lemme 3.7 pouvant être borné sans faire appel à* $n_{\bar{B}}$, m *ou* n.

La constante A ne dépend pas de n_B, $n_{\bar{B}}$ ou n et sa valeur explicite peut être trouvée dans le théorème 3.5. Le terme dominant dans le le dénominateur est $A\mathbb{V}_\infty^2 n_{\bar{B}}^3 n_{\mathscr{B}}^3 \kappa!^2\lceil\ln n\rceil^\kappa n$. C'est $\mathcal{O}(n(\ln n)^\kappa)$, alors que pour l'inégalité de Hoeffding i.i.d., ce terme est $\mathcal{O}(n)$.
De plus, ce terme est fortement impacté par la dimension du champ aléatoire κ. Cependant, cette limitation n'est pas un problème en pratique, car, dans la plupart des cas pratiques, $\kappa \in 1,2$.
D'autres facteurs importants sont les dimensions paramétriques de notre modèle, représentées par n_B et $n_{\bar{B}}$. Assez logiquement, la qualité de l'inégalité diminue lorsque le nombre de variables à contrôler augmente. Notre inégalité est assez sensible à ces facteurs et n'est donc pas un résultat approprié pour une estimation de grande dimension, c'est-à-dire lorsque n est inférieur à n_B et $n_{\bar{B}}$. Rappelons que pour une chaîne de Markov d'ordre 1, ces deux termes seraient égaux à 2.

Enfin, la condition $\varepsilon \geq 2n_{\bar{B}}\mathbb{V}_\infty$ ne semble pas restrictive pour les applications, comme l'exemple d'inégalité oracle de la sous-section 3.3.2 le montre .

Énonçons maintenant d'une version simplifiée de l'inégalité de concentration pour les hypothèses les plus faibles $(\mathbf{H_1^m})$ et $(\mathbf{H_2^m})$

Théorème 3.2. INÉGALITÉ TYPE-HOEFFDING, CADRE NON CAUSAL, CONTRACTION FAIBLE *Soit* $n = Card(\mathscr{I})$, $n_B = Card(\mathscr{B}) + 1$ *et* $n_{\bar{B}} = Card(\bar{\mathscr{B}})$. *Si* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et* $(\mathbf{H_4})$ *sont vérifiés, alors il existe des constantes* A,B,C,D,E,F,H *telles que, pour* $\varepsilon \geq 2F\left(n_{\bar{B}}n_B\right)^3\kappa!\lceil\ln(n)\rceil^{2\kappa}n^{\frac{2}{m}}$.

$$\mathbb{P}\left(|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\left(\frac{\varepsilon}{2} - F\left(n_{\bar{B}}n_B\right)^3\kappa!\lceil\ln(n)\rceil^{2\kappa}n^{\frac{2}{m}}\right)^2}{\left(Hn_{\bar{B}}n^{\frac{2}{m}}\right)^2\left(1 + En_{\bar{B}}n_B^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right)$$
$$+\frac{\rho^m}{n}\left(2n_Bn_{\bar{B}}C\lceil\ln(n)\rceil^\kappa + \left(\frac{D}{n_B^3n_{\bar{B}}^2\Upsilon(d)\ln(n)^{2\kappa}}\right)^m\right).$$

On peut détailler ces constantes :

- $A = \frac{1+\frac{\ln(\rho^{-1})}{\kappa!n_B} + \ln(\rho^{-1})\frac{\left(\frac{\kappa-1}{e}\right)^{\kappa-1}}{(\kappa-1)!}}{\ln(\rho^{-1})^\kappa}$, *donc* $\Upsilon(d) \leq A\kappa!$, *où* Υ *est la fonction définie au Lemme 3.7 pouvant être bornée sans faire appel à* $n_{\bar{B}}$, m *ou* n.
- $B = \frac{1}{\rho^2} + \frac{1}{\rho\left(n_{\bar{B}}n_B\right)^{\frac{1}{m}}} + \frac{2M}{\Upsilon(d)\mathbb{V}_m\left(n_{\bar{B}}n_B\right)^2}$.
- $C = \left\lceil \frac{1-\frac{1}{m}}{\ln(\rho^{-1})} \right\rceil^\kappa$, *donc* $d^\kappa \leq C\lceil\ln(n)\rceil^\kappa$.
- $D = \frac{\rho\ln(\rho^{-1})^{2\kappa}}{2\left(1-\frac{1}{m}\right)^{2\kappa}}$.
- E *tel que* $\Upsilon(d)^2 d^\kappa \leq E(\kappa!)^2\lceil\ln(n)\rceil^\kappa$.

— $F = 2BC^2A\mathbb{V}_m$.

— $H = \frac{\mathbb{V}_m}{\rho}$

Des commentaires similaires sur la dimension κ et les paramètres n_B, $n_{\bar{B}}$ à ceux concernant le théorème précédent peuvent être faits.
Le dénominateur dans l'exponentielle est $\mathcal{O}\left(nn^{\frac{4}{m}}(\ln(n))^{\kappa}\right)$. Par rapport au cas i.i.d, il y a un terme supplémentaire dans $n^{\frac{4}{m}}(\ln(n))^{\kappa}$ que nous devrions expliquer. Le terme $n^{\frac{4}{m}}$ vient de l'hypothèse $(\mathbf{H_1^m})$ qui est moins forte que l'hypothèse classique de contraction. A cause de ce terme, l'équation n'est intéressante que si $m > 4$. Comme dans le cas de la contraction forte, le terme $(\ln(n))^{\kappa}$ provient de la dépendance, et est obtenu par article similaire pour la dépendance unidirectionnelle. [DF15].

$(\mathbf{H_1^m})$ et $(\mathbf{H_2^m})$ sont des hypothèses assez faibles que $(\mathbf{H_1^\infty})$ et $(\mathbf{H_2^\infty})$ et conduisent ainsi à des inégalités de concentration dégradées. En effet, au terme exponentiel, le dénominateur est asymptotiquement dominé par $\mathcal{O}\left(n^{1+\frac{4}{m}}(\ln(n))^{\kappa}\right)$ au lieu de $\mathcal{O}\left(n(\ln(n))^{\kappa}\right)$ sous les hypothèses $(\mathbf{H_1^\infty})$ et $(\mathbf{H_2^\infty})$ et $\mathcal{O}(n)$ dans le cas i.i.d. .

Les deux termes additifs supplémentaires diminuent plus rapidement que le terme principal et ne sont pas dominants pour les applications comme nous le montrerons dans la sous-section 3.3.2.

Ces deux théorèmes conduisent aux corollaires suivants pour l'espérance de la déviation lorsque qu'on les intègre.!

Corollaire 3.1. BORNE SUR L'ESPÉRANCE DE DEVIATION, CAS UNIFORME *On suppose que* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés. Soit* A *la constante définie au Théorème 3.1. On a la relation :*

$$\mathbb{E}[|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]|] \leq n_{\bar{B}}\mathbb{V}_\infty\left(2 + \sqrt{\frac{\pi}{2}\left(1 + An_{\bar{B}}n_B^3(\kappa!)^2\lceil\ln(n)\rceil^{\kappa}n\right)}\right).$$

Corollaire 3.2. BORNE SUR L'ESPÉRANCE DE DEVIATION, CAS FAIBLE *On suppose que* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *and* $(\mathbf{H_4})$ *sont verifiés. Soit* A,B,C,D,E,F,H *les constantes définies au théorème 3.2. On a la relation :*

$$\begin{aligned}
&\mathbb{E}[|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]|] \\
&\leq 2n_{\bar{B}}Hn^{\frac{2}{m}}\sqrt{\frac{\pi}{2}\left(1 + En_{\bar{B}}n_B^3(\kappa!)^2\lceil\ln(n)\rceil^{\kappa}n\right)} + 2F\left(n_{\bar{B}}n_B\right)^3\kappa!\lceil\ln(n)\rceil^{2\kappa}n^{\frac{2}{m}} \\
&+ 2\rho^m M\left(2n_B n_{\bar{B}}C\lceil\ln(n)\rceil^{\kappa} + \left(\frac{D}{n_B^3 n_{\bar{B}}^2\Upsilon(d)\ln(n)^{2\kappa}}\right)^m\right).
\end{aligned}$$

Remarque 3.3. *D'autres types d'inégalités de concentration (Bernstein, von Bahr-Esseen) auraient pu être prouvés dans le même contexte en utilisant la même approche. En effet, il devrait aussi être possible, comme dans l'i.i.d. cas, pour dériver des inégalités supplémentaires à partir de l'inégalité exponentielle énoncée dans le lemme 3.9.*

Néanmoins, nous nous concentrons sur l'inégalité de Hoeffding pour sa simplicité et son utilisation dans de nombreuses applications.

3.3.2 Inégalité Oracle pour le problème de complétion

Dans cette sous-section, on propose une application de nos résultats à la théorie de l'apprentissage. On adapte l'approche de [BBL00], [Lin02] et [Lug02a] dans le cadre non-causal, pour obtenir un oracle lié au problème de sélection de modèle similaire à celle présentée en section 1.2.2.

Application au problème de complétion

Un problème classique en théorie de l'apprentissage est le problème d'achèvement (à la fois en régression ou en classification). L'objectif est de prédire X_t en utilisant ses voisins sur un réseau dimensionnel κ $(X_t)_{t\in\bar{\mathscr{B}}}$. La prédiction est donnée par un modèle $\hat{f}$

$$\widehat{x}_s = \hat{f}((x_{s+t})_{t\in\bar{\mathscr{B}}\setminus\{0\}}).$$

Ce problème de complétion peut sembler trivial et de peu d'intéret, mais est en fait à la base de nombreuses tâches en NLP. En effet, la prédiction de mots est souvent utilisée comme tâche principale pour entraîner le codeur dans un schéma codeur-décodeur. Il crée un modèle de langage, c'est-à-dire une distribution de probabilité d'un mot dans un texte. Ce modèle de langage est ensuite utilisé pour des tâches plus complexes (traduction, segmentation de texte, réponse aux questions, etc.).

Comme aux chapitres 1 et 2, on introduit aussi une fonction de perte $L : \mathscr{X}^2 \mapsto \mathbb{R}$. $L(\widehat{x}_s, x_s)$ quantifie l'erreur commise lorsque l'algorithme considéré prédit $\widehat{x}_s$ au lieu de x_s. Dans ce cas, la fonction d'intérêt Φ introduite précédemment correspond à la perte de $\hat{f}$ sur un échantillon

$$\Phi((X_{s+t})_{t\in\bar{\mathscr{B}}}) = L(\hat{f}((X_{s+t})_{t\in\bar{\mathscr{B}}\setminus\{0\}}), X_s).$$

et

$$S_{\mathscr{I}} = \sum_{s\in\mathscr{I}} \Phi((X_{s+t})_{t\in\bar{\mathscr{B}}}) = \sum_{s\in\mathscr{I}} L(\hat{f}((X_{s+t})_{t\in\bar{\mathscr{B}}\setminus\{0\}}), X_s).$$

Remarque 3.4. *On insiste le fait que $\mathscr{B}$ et $\bar{\mathscr{B}}$ peuvent être différents en pratique. En effet, en pratique, on ne peut pas connaître la taille de $\mathscr{B}$. On peut également également vouloir essayer un modèle plus simple (avec $n_{\bar{\mathscr{B}}} < n_{\mathscr{B}}$) pour réduire le sur-apprentissage et augmenter la taille de l'ensemble d'entrainement. On peut voir cette effet sur la figure 3.1, où $\bar{\mathscr{B}}$ (en rouge) est plus petit que $\mathscr{B}$.*

Remarque 3.5. *La fonction de perte $L : \mathscr{X}^2 \mapsto \mathbb{R}$ peut être choisie par l'utilisateur. Les choix classiques incluent l'entropie croisée, la divergence de Kullback-Leibler, etc. Lorsque $\mathscr{X}$ n'est pas borné, certaines fonctions de coût classiques telles que la distance quadratique doivent être tronquées pour satisfaire* $(\mathbf{H_4})$.

Remarque 3.6. *Si* $(\mathbf{H_4})$ *est vérifié avec une borne supérieure donnée* M, *il est toujours possible de borner cette fonction de coût par une constante puis de diviser cette fonction par cette constante. Par conséquent, on peut toujours supposer que* $M = 1$ *dans l'hypothèse* $(\mathbf{H_4})$ *sans perte de généralité.*

Sur la Figure 3.1, le voisinage rouge $\mathscr{V}(\bar{\delta}, t)$ est utilisé pour calculer $\widehat{x}_t = \hat{f}((x_{t+s})_{s\in\bar{\mathscr{B}}})$. Cependant, en pratique, pour effectuer ce calcul, il faut connaître tous les points de ce voisinage $\mathscr{V}(\bar{\delta}, t)$. Cela implique donc que l'on ne peut pas utiliser tous les points connus dans l'ensemble d'entraînement $\mathscr{I}$, seulement ceux qui ne sont pas trop loin du bord.

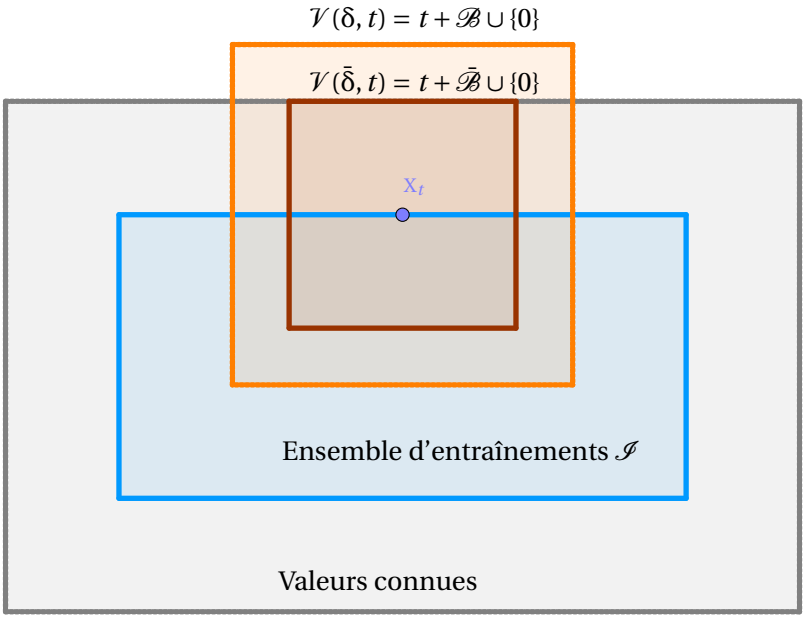

FIGURE 3.1 – Principe de l'algorithme de complétion

Le risque empirique et théorique sont définies exactement comme aux chapitres précédent.

$$\mathbb{L}(\hat{f}) = \mathbb{E}[\mathrm{L}(\hat{f}((\mathrm{X}_{s+t})_{t\in\bar{\mathcal{B}}\setminus\{0\}}), \mathrm{X}_s)], \tag{3.8}$$

$$\widehat{\mathrm{L}}_n(\hat{f}) = \frac{1}{n}\sum_{s\in\mathcal{I}} \mathrm{L}(\hat{f}((\mathrm{X}_{s+t})_{t\in\bar{\mathcal{B}}\setminus\{0\}}), \mathrm{X}_s). \tag{3.9}$$

L'application immédiate de nos résultats donne des inégalités de déviation pour $|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|$ sous les hypothèses $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ et $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$.

Théorème 3.3. INÉGALITÉS DE DÉVIATION : *Soit* $n_{\mathrm{B}} = Card(\mathcal{B}) + 1$ *et* $n_{\bar{\mathrm{B}}} = Card(\bar{\mathcal{B}})$.

— *Si* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés,alors il y a une constante définie comme au théorème 3.1, telle que pour* $\varepsilon > \frac{2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}$, *on ait :*

$$\mathbb{P}\left(|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2n^2\left(\varepsilon - \frac{2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right).$$

— *Si* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et* $(\mathbf{H_4})$ *sont vérifiés, alors il y a des constantes définies au théorème 3.2 telles que pour* $\varepsilon \geq \frac{2\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}$, *on ait :*

$$\mathbb{P}\left(|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2n^{2-\frac{4}{m}}\left(\frac{\varepsilon}{2} - \frac{\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)^2}{\left(\mathrm{H}n_{\bar{\mathrm{B}}}\right)^2\left(1 + \mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right) + \frac{\rho^m}{n}\mathrm{L}_2(n),$$

avec $\mathrm{L}_1(n) = \mathrm{F}\left(n_{\bar{\mathrm{B}}}n_{\mathrm{B}}\right)^3\kappa!\lceil\ln(n)\rceil^{2\kappa}$ *et* $\mathrm{L}_2(n) = 2n_{\mathrm{B}}n_{\bar{\mathrm{B}}}\mathrm{C}\lceil\ln(n)\rceil^\kappa + \left(\frac{\mathrm{D}}{n_{\mathrm{B}}^3 n_{\bar{\mathrm{B}}}^2\Upsilon(d)\ln(n)^{2\kappa}}\right)^m$.

Démonstration. La preuve est une application directe de 3.1 et 3.2 avec $\mathrm{S}_{\mathcal{I}} = \sum_{s\in\mathcal{I}} \mathrm{L}(\hat{f}((\mathrm{X}_{s+t})_{t\in\bar{\mathcal{B}}\setminus\{0\}}), \mathrm{X}_s)$. □

Si $\mathbb{E}[\exp(|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|)]$ existe, alors on a le très utile corollaire suivant .

Corollaire 3.3.

— *Si* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés,alors pour tout* $s > 0$,

$$\mathbb{E}[\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|)] \leq \exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n} + \frac{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}{8n^2}\right)$$
$$\times\left(1 + \frac{n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s}{n}\sqrt{2\pi\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right). \tag{3.10}$$

— *Si* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et* $(\mathbf{H_4})$ *sont vérifiés, alors pour tout* $s > 0$,

$$\begin{aligned}&\mathbb{E}[\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|)]\\ &\leq \exp\left(\frac{1}{2n^{1-\frac{4}{m}}}\left(4s\mathrm{L}(n) + \left(\mathrm{H}n_{\bar{\mathrm{B}}}s\right)^2\left(1 + \mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)\right)\right)\\ &\times\left(1 + \frac{\exp(\mathrm{M}s)\rho^m\mathrm{L}_2(n)}{n} + \frac{2\mathrm{H}n_{\bar{\mathrm{B}}}s}{n^{1-\frac{2}{m}}}\sqrt{2\pi\left(1 + \mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right).\end{aligned} \tag{3.11}$$

La preuve de ce résultat est dans l'appendice B.1.

Sélection de modèles

Dans cette sous-section, on étend le cadre classique de la sélection de modèles en se servant des notations du chapitre 1.

Comme en section 1.1, on considère un ensemble fini S du modèle et on s'intéresse au minimiseur du risque empirique $\tilde{f} = \underset{\hat{f}\in\mathrm{S}}{\operatorname{argmin}}\left(\widehat{\mathrm{L}}_n(\hat{f})\right)$ et au meilleur modèle théorique sur S, $f_\mathrm{S}^* = \underset{\hat{f}\in\mathrm{S}}{\operatorname{argmin}}\left(\mathbb{L}(\hat{f})\right)$. On cherche à majorer l'erreur d'estimation $\mathbb{L}(\tilde{g}) - \mathbb{L}(g_\mathrm{S}^*)$. On rappelle également que N est le cardinal de S.

Corollaire 3.4. DÉVIATION DE L'ERREUR D'ESTIMATION *Si* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés, alors pour* $\varepsilon > \frac{4n_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}$,

$$\mathbb{P}\left(\mathbb{L}(\tilde{f}) - \mathbb{L}(f_\mathrm{S}^*) > \varepsilon\right) \leq 2\mathrm{N}\exp\left(\frac{-2n^2\left(\frac{\varepsilon}{2} - \frac{2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right). \tag{3.12}$$

Si $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et* $(\mathbf{H_4})$ *sont vérifiés, alors pour* $\varepsilon \geq \frac{4\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}$,

$$\mathbb{P}\left(\mathbb{L}(\tilde{f}) - \mathbb{L}(f_\mathrm{S}^*) > \varepsilon\right) \leq \mathrm{N}\left(2\exp\left(\frac{-2n^{2-\frac{4}{m}}\left(\frac{\varepsilon}{4} - \frac{\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)^2}{\left(\mathrm{H}n_{\bar{\mathrm{B}}}\right)^2\left(1 + \mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right) + \frac{\rho^m}{n}\mathrm{L}_2(n)\right). \tag{3.13}$$

Démonstration.

$$\begin{aligned}\mathbb{L}(\tilde{f}) - \mathbb{L}(f_\mathrm{S}^*) &= \mathbb{L}(\tilde{f}) - \widehat{\mathrm{L}}_n(\tilde{f}) + \widehat{\mathrm{L}}_n(\tilde{f}) - \widehat{\mathrm{L}}_n(f_\mathrm{S}^*) + \widehat{\mathrm{L}}_n(f_\mathrm{S}^*) - \mathbb{L}(f_\mathrm{S}^*)\\ &\leq \left(\mathbb{L}(\tilde{f}) - \widehat{\mathrm{L}}_n(\tilde{f})\right) + \left(\widehat{\mathrm{L}}_n(\tilde{f}) - \widehat{\mathrm{L}}_n(f_\mathrm{S}^*)\right) + \left(\widehat{\mathrm{L}}_n(f_\mathrm{S}^*) - \mathbb{L}(f_\mathrm{S}^*)\right)\\ &\leq 2\sup_{\hat{f}\in\mathrm{S}}|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})| + \left(\widehat{\mathrm{L}}_n(\tilde{f}) - \widehat{\mathrm{L}}_n(f_\mathrm{S}^*)\right)\\ &\leq 2\sup_{\hat{f}\in\mathrm{S}}|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})| \quad \text{car } \left(\widehat{\mathrm{L}}_n(\tilde{f}) - \widehat{\mathrm{L}}_n(f_\mathrm{S}^*)\right) \leq 0.\end{aligned}$$

Donc , pour tout ε,

$$\begin{aligned}\mathbb{P}\left(\mathbb{L}(\tilde{f})-\mathbb{L}(f_S^*)>\varepsilon\right) &\leq \mathbb{P}\left(\sup_{\hat{f}\in S}|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|>\frac{\varepsilon}{2}\right)\\ &\leq \sum_{\hat{f}\in S}\mathbb{P}\left(|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|>\frac{\varepsilon}{2}\right) \quad \text{par borne d'union}\\ &\leq N\mathbb{P}\left(|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|>\frac{\varepsilon}{2}\right).\end{aligned}$$

Ensuite, le théorème 3.3 donne le résultat pour les deux hypothèses. □

On peut également obtenir des bornes en esperance. On s'intéresse d'abord au cadre de la contraction uniforme.

Corollaire 3.5. INÉGALITÉ ORACLE POUR L'ERREUR D'ESTIMATION *Si* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés, alors*

$$\mathbb{E}[\mathbb{L}(\tilde{f})-\mathbb{L}(f_S^*)]\leq\frac{2n_{\bar{B}}\mathbb{V}_\infty}{n}\left(2+\sqrt{\left(1+An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\left(\sqrt{\frac{\ln(N)}{2}}+\sqrt{2\pi}\right)\right).$$

Asymptotiquement, cela donne la relation suivante :

$$\mathbb{E}[\mathbb{L}(\tilde{f})-\mathbb{L}(f_S^*)]\leq H(n_{\bar{B}},n_B,\kappa,\rho,N,n)\underset{n\to\infty}{\sim}2\mathbb{V}_\infty\left(\frac{\sqrt{\ln(N)}}{2}+\sqrt{2\pi}\right)n_{\bar{B}}n_B\kappa!\sqrt{An_{\bar{B}}n_B}\sqrt{\frac{\ln(n)^\kappa}{n}}.$$

Démonstration. On a montré dans le corollaire précédent que $\mathbb{L}(\tilde{f})-\mathbb{L}(f_S^*)\leq 2\sup_{\hat{f}\in S}|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|$.

De plus, pour tout $s>0$,

$$\begin{aligned}\mathbb{E}[\sup_{\hat{f}\in S}|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|] &= \frac{1}{s}\mathbb{E}\left[\ln\left(\sup_{\hat{f}\in S}\exp\left(s|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|\right)\right)\right]\\ &\leq \frac{1}{s}\ln\left(\mathbb{E}\left[\sup_{\hat{f}\in S}\exp\left(s|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|\right)\right]\right) \quad \text{using Jensen inequality}\\ &\leq \frac{1}{s}\ln\left(\mathbb{E}\left[\sum_{\hat{f}\in S}\exp\left(s|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|\right)\right]\right)\\ &= \frac{1}{s}\ln\left(N\mathbb{E}\left[\exp\left(s|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|\right)\right]\right).\end{aligned}$$

Si $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ et $(\mathbf{H_3})$ sont vérifiés, en utilisant l'équation (3.10) du corollaire 3.3, on obtient

$$\begin{aligned}\mathbb{E}[\mathbb{L}(\tilde{f})-\mathbb{L}(f_S^*)]\leq&\frac{2\ln(N)}{s}+\frac{4n_{\bar{B}}\mathbb{V}_\infty}{n}+\frac{s\left(n_{\bar{B}}\mathbb{V}_\infty\right)^2\left(1+An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}{4n^2}\\ &+\frac{2n_{\bar{B}}\mathbb{V}_\infty}{n}\sqrt{2\pi\left(1+An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}, \text{ car } \ln(1+x)\leq x.\end{aligned}$$

En choisissant $s=\dfrac{2n}{n_{\bar{B}}\mathbb{V}_\infty}\sqrt{\dfrac{2\ln(N)}{1+An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n}}$, on obtient

$$\begin{aligned}&\mathbb{E}[\mathbb{L}(\tilde{f})-\mathbb{L}(f_S^*)]\\ &\leq\frac{2n_{\bar{B}}\mathbb{V}_\infty}{n}\left(\sqrt{\frac{\ln(N)}{2}\left(1+An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}+2+\sqrt{2\pi\left(1+An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right)\\ &\leq\frac{2n_{\bar{B}}\mathbb{V}_\infty}{n}\left(2+\sqrt{\left(1+An_{\bar{B}}n_B^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\left(\sqrt{\frac{\ln(N)}{2}}+\sqrt{2\pi}\right)\right).\end{aligned}$$

□

Remarque 3.7. *Dans le cas* $\mathbb{E}\left[\exp\left(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|\right)\right]$ *n'existe pas, ou si l'on a seulement* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *and* $(\mathbf{H_4})$, *on peut utiliser une borne plus simple en remarquant que*

$$\mathbb{E}[\mathbb{L}(\tilde{f}) - \mathbb{L}(f_{\mathrm{S}}^*)] \leq \mathbb{E}[\sup_{\hat{f}\in \mathrm{S}} |\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|] \leq \mathrm{N}\mathbb{E}[|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|]$$

et utilisant les corollaires 3.1 ou 3.2. En faisant ainsi, la vitesse d'apprentissage N *à la place de* $\sqrt{\ln(\mathrm{N})}$. *Si l'on suppose que* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et* $(\mathbf{H_4})$,*on obtient la borne asymptotique suivante*

$$\mathbb{E}[\mathbb{L}(\tilde{f}) - \mathbb{L}(f_{\mathrm{S}}^*)] \leq \mathrm{H}(n_{\bar{\mathrm{B}}}, n_{\mathscr{B}}, \kappa, \rho, \mathrm{N}, n) \underset{n\to\infty}{\sim} \frac{\mathrm{H}\kappa!\sqrt{2\mathrm{E}\pi(n_{\bar{\mathrm{B}}} n_{\mathrm{B}})^3 \ln(n)^\kappa}}{n^{\frac{1}{2}-\frac{2}{m}}}.$$

Où H et E sont les constantes définies dans le corollaire 3.2.

Remarque 3.8. *Dans [Lug02a], une borne est fournie pour la déviation maximale attendue dans le cadre indépendant :*

$$\mathbb{E}[\sup_{\hat{f}\in \mathrm{S}} |\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|] \leq \sqrt{\frac{\ln(2\mathrm{N})}{2n}}.$$

Dans notre cadre, avec les hypothèses $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$, *nous avons obtenu des bornes comparables (Corollaire 3.5) avec un terme supplémentaire* $\sqrt{\ln(n)^\kappa}$.

3.4 Approximation de champ non-causal

Cette section, et la suivante sont consacrées à la preuve des résultats de la section précédente 3.3. Cette section traite plus précisément de l'introduction d'une approximation du champ de vecteur (X_t) par un nouveau champ $(\tilde{\mathrm{X}}_t^{[d]})$ telle que chaque variable $\tilde{\mathrm{X}}_t^{[d]}$ soit une fonction $\tilde{\mathrm{H}}^{[d]}$ d'un nombre *fini* de variables aléatoires. Cette approche est analogue à celle introduite par [CW18] pour des résultats asymptotiques dans des champs de vecteurs non causaux.

3.4.1 Approximation de X_t

Dans cette section, on introduit $\tilde{\mathrm{X}}_t^{[d]}$ et on contrôle son écart avec X_t.

Notations

Rappelons et introduisons d'abord quelques notations pour gérer les champs aléatoires.

— On a déjà introduit la notation $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t\in\mathbb{Z}^\kappa}$. De même, on utilise la notation $\boldsymbol{\varepsilon}' = (\varepsilon'_t)_{t\in\mathbb{Z}^\kappa}$ où ε'_t sont aussi des variables aléatoires $\Omega \mapsto \mathrm{E}$.

— Pour $s \in \mathbb{Z}^\kappa$, θ_s est l'opérateur de décalage de s, c'est à dire que pour $s \in \mathbb{Z}^\kappa$, $\theta_s((\varepsilon_t)_{t\in\mathbb{Z}^\kappa}) = (\varepsilon_{s+t})_{t\in\mathbb{Z}^\kappa}$.

— Pour $s \in \mathbb{Z}^\kappa$, on note $\boldsymbol{\varepsilon}_s$ le champ de de vecteurs décalé par θ_s, c'est à dire le champ de vecteurs $\boldsymbol{\varepsilon}_s = \theta_s(\boldsymbol{\varepsilon}) = \theta_s((\varepsilon_t)_{t\in\mathbb{Z}^\kappa}) = (\varepsilon_{s+t})_{t\in\mathbb{Z}^\kappa}$.

Reconstruction exacte

Le théorème 1 de [DT07] assure, dans des conditions convenables sur F, l'existence et l'unicité d'une fonction H telle que, pour chaque $t \in \mathbb{Z}^\kappa$

$$\mathrm{X}_t = \mathrm{H}(\boldsymbol{\varepsilon}_t). \tag{3.14}$$

On peut faire deux commentaires par rapport à ce résultat

— Ce théorème fournit une expression de chaque X_t selon un nombre infini de i.i.d. variables aléatoires (tout le champ aléatoire $\boldsymbol{\varepsilon}$). Cependant, de nombreuses inégalités de concentration n'impliquent qu'un nombre fini de variables aléatoires.

— Ce théorème repose sur une hypothèse de contraction absolue sur F (similaire à l'équation 3.3) qui est une hypothèse plus forte que $(\mathbf{H}_1^{\infty})$ et $(\mathbf{H}_1^{\mathbf{m}})$.

Pour ces deux raisons, on ne peut pas utiliser cette reconstruction *exacte* de la solution X_t de l'équation (3.2) pour établir des inégalités de concentration sur des champs de vecteurs non-causaux en utilisant les hypothèse introduites plus tôt.

Intuition

Essayons de donner une intuition derrière l'approximation que l'on va allons introduire. L'idée principale est d'approcher chaque X_t par une autre variable aléatoire $\tilde{X}_t$ qui, comme $H(\boldsymbol{\varepsilon})$, dépend de l'innovation $\boldsymbol{\varepsilon}$.

Cependant, contrairement à $H(\boldsymbol{\varepsilon})$, on n'utilisera qu'un nombre fini de variable aléatoire ε_t, c'est à dire celles qui sont situées dans un voisinage fini autour de X_t.

Définition de l'approximation $\tilde{H}^{[d]}$

Rappelons quelques notations sur les voisinages, et la notion de dilatation d'un tel voisinage

— Dans l'équation $\forall t \in \mathbb{Z}^{\kappa}, X_t = F((X_{t+s})_{s\in\mathscr{B}}, \varepsilon_t)$ (Équation (3.2)), $\mathscr{B}$ est un $\kappa-$orthotope défini par $\mathscr{B} = \mathscr{V}(\delta, 0)\setminus\{0\} = \prod_{i=1}^{\kappa}[-\delta_i, \delta_i]\setminus\{0\}$.

— On introduit la dilatation d'un tel orthotope$\kappa-$ orthotope $\mathscr{V}(\delta, 0)$ par un facteur d comme le nouvel orthotope $\mathscr{V}(d\delta, s)$ centré sur le même point s tel que :

$$\mathscr{V}(d\delta, s) = \{t = (t_1, \ldots, t_{\kappa}) \in (\mathbb{Z})^{\kappa} / \forall i, s_i - d\delta_i \leq t_i \leq s_i + d\delta_i\}. \tag{3.15}$$

On va se servir de cette notion pour définir formellement la fonction $\tilde{H}^{[d]}$.

Définition 3.7. FONCTION D'APPROXIMATION $\tilde{H}^{[d]}$ *On définit récursivement la fonction* $\tilde{H}^{[d]}$ *par*

$$\tilde{H}^{[d]}(X, (\varepsilon_t)_{t\in\mathscr{V}(d\delta,s)}) = \begin{cases} X \text{ si } d = 0, \\ F\left((\tilde{H}^{[d-1]}\left(X, (\varepsilon_u)_{u\in\mathscr{V}(\delta(d-1),t+s)}\right))_{s\in\mathscr{B}}, \varepsilon_s\right) \text{ sinon.} \end{cases} \tag{3.16}$$

Finalement, nous sommes en mesure de définir l'approximation $\tilde{X}_t^{[d]}$ à l'aide de la fonction précédente.

Définition 3.8. APPROXIMATION $\tilde{X}_t^{[d]}$

$$\forall d \in \mathbb{N}, \forall t \in \mathbb{Z}^{\kappa}, \tilde{X}_t^{[d]} = \tilde{H}^{[d]}(\tilde{X}, (\varepsilon_s)_{s\in\mathscr{V}(d\delta,t)}). \tag{3.17}$$

Où $\tilde{X}$ *est une variable aléatoire indépendante suivant la distribution* μ_X.

On peut reformuler l'équation (3.16) en utilisant la notation $\tilde{X}_t^{[d]}$.

$$\forall d \in \mathbb{N}, \forall t \in \mathbb{Z}^{\kappa}, \tilde{X}_t^{[d]} = \begin{cases} \tilde{X} \text{ if } d = 0, \\ F\left((\tilde{X}_{t+s}^{[d-1]})_{s\in\mathscr{B}}, \varepsilon_t\right) \text{ else.} \end{cases}$$

Ainsi, $\tilde{X}_t^{[d]}$ est une approximation de X_t impliquant des variables aléatoires ε_t qui appartiennent au voisinage fini $\mathscr{V}(d\delta, t)$. En dehors de ce voisinage, on complète l'approximation avec une variable aléatoire $\tilde{X}$ tirée de la loi μ_X et indépendante de $(X_t)_{t\in\mathbb{Z}^{\kappa}}$ et $\boldsymbol{\varepsilon}$.

Remarque 3.9. *Soulignons que si la fonction H du théorème 1 de [DT07] existe et est unique, alors* $\lim_{d\to\infty} \tilde{H}^{[d]}(X, (\varepsilon_t)_{t\in\mathscr{V}(d\delta,t)}) = H(\boldsymbol{\varepsilon})$.
Néanmoins cette limite peut ne pas exister et ne pas être unique, mais même dans ce cas, pour tout d fini, l'approximation $\tilde{H}^{[d]}(X, (\varepsilon_t)_{t\in\mathscr{V}(d\delta,t)})$ *est toujours définie.*

Erreur d'approximation $\|X_t - \tilde{X}_t^{[d]}\|_m$

L'approximation $\tilde{X}_t^{[d]}$ n'est utile que si nous sommes capables de contrôler l'erreur d'approximation en norme m $\|X_t - \tilde{X}_t^{[d]}\|_m$. C'est l'objet des deux lemmes suivants.

Lemme 3.2. CONTRÔLE DE L'ERREUR D'APPROXIMATION *Soit* $m \in \mathbb{N} \cup \{\infty\}$. *Si* $(\mathbf{H_1^m})$ *et* $(\mathbf{H_2^m})$ *sont vérifiées,alors :*

$$\forall t \in \mathbb{Z}^\kappa, \forall d \in \mathbb{N}, \|X_t - \tilde{X}_t^{[d]}\|_m \leq \rho^d \mathbb{V}_m.$$

La preuve peut être trouvée dans l'annexe B.2. Ce lemme est la clé pour contrôler la qualité de l'approximation et est à la base des différents résultats du chapitre. En particulier, il explique la forme un peu particulière de l'hypothèse $(\mathbf{H_1^m})$ et la nécessité de la formuler "pour tout couplage".

Statistique approchée $\tilde{S}_{\mathscr{J}}^{[d]}$

Précédemment, on définissait la statistique $S_{\mathscr{J}}$ (voir l'équation (3.5)), en tant que fonction des variables aléatoires $(X_t)_{t\in\mathscr{J}}$. Nous introduisons maintenant un équivalent de la statistique $S_{\mathscr{J}}$ qu'on appellera $\tilde{S}_{\mathscr{J}}^{[d]}$ en tant que fonction des variables aléatoires $(\tilde{X}_t^{[d]})_{t\in\mathscr{J}}$.

On rappelle :

$$S_{\mathscr{J}} = \sum_{t\in\mathscr{J}} \Phi((X_{t+s})_{s\in\bar{\mathscr{B}}}). \tag{3.18}$$

De manière similaire, on définit :

$$\tilde{S}_{\mathscr{J}}^{[d]} = \sum_{t\in\mathscr{J}} \Phi((\tilde{X}_{t+s}^{[d]})_{s\in\bar{\mathscr{B}}}). \tag{3.19}$$

En utilisant le lemme 3.2, on peut contrôler la différence entre $S_{\mathscr{J}}$ et $\tilde{S}_{\mathscr{J}}^{[d]}$. C'est l'objet du corollaire suivant.

Corollaire 3.6. CONTRÔLE DU MOMENT DE L'ÉCART $|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}|$ *Soit* $m \in \mathbb{N} \cup \{\infty\}$. *Si* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$ *et* $(\mathbf{H_3})$ *sont vérifiées,alors en rappelant que par définition* $n = Card(\mathrm{I})$ *and* $n_{\bar{\mathrm{B}}} = Card(\bar{\mathrm{B}})$ *:*

$$\|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}\|_m \leq n n_{\bar{\mathrm{B}}} \rho^d \mathbb{V}_m. \tag{3.20}$$

Démonstration. Par définition de $S_{\mathscr{J}}$ et $\tilde{S}_{\mathscr{J}}^{[d]}$ (Équations (3.18) and (3.19)), en utilisant $(\mathbf{H_3})$, on a :

$$\|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}\|_m \leq \sum_{t\in\mathscr{J}} \|\Phi((X_{t+s})_{s\in\bar{\mathscr{B}}}) - \Phi((\tilde{X}_{t+s}^{[d]})_{s\in\bar{\mathscr{B}}})\|_m \leq \sum_{t\in\mathscr{J}} \sum_{s\in\bar{\mathscr{B}}} \|X_{t+s} - \tilde{X}_{t+s}^{[d]}\|_m.$$

Et grâce au lemme 3.2

$$\|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}\|_m \leq n n_{\bar{\mathrm{B}}} \rho^d \mathbb{V}_m.$$

□

3.4.2 Inégalité de concentration pour $\tilde{S}_{\mathscr{J}}^{[d]}$

Dans cette sous-section, nous allons établir une inégalité de concentration pour $\tilde{S}_{\mathscr{J}}^{[d]}$. En effet, $\tilde{S}_{\mathscr{J}}^{[d]}$ peut être vu comme une fonction d'un nombre fini de variables aléatoires, il est donc possible d'établir une inégalité de McDiarmid pour $\tilde{S}_{\mathscr{J}}^{[d]}$.

Plus précisément, nous visons à montrer le théorème suivant.

Théorème 3.4. INÉGALITÉ DE CONCENTRATION POUR $\tilde{S}_{\mathscr{I}}^{[d]}$ *Soit* $d \in \mathbb{N}$. *Si l'on suppose que* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiées, alors*

$$\forall \varepsilon > 0, \mathbb{P}\left(|\tilde{S}_{\mathscr{I}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{I}}^{[d]}\right]| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\varepsilon^2}{\left(n_{\bar{B}}\mathbb{V}_\infty\right)^2\left(n^2\rho^{2d} + N_2\left(n_B\Upsilon(d)\right)^2\right)}\right). \tag{3.21}$$

Si l'on suppose que $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et* $(\mathbf{H_4})$ *sont vérifiées, alors* $\forall (t_1, t_2) > 0$,

$$\forall \varepsilon > 2npM + p\bar{c}, \mathbb{P}\left(|\tilde{S}_{\mathscr{I}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{I}}^{[d]}\right]| \geq \varepsilon\right) \leq 2\left(p + \exp\left(\frac{-2(\varepsilon - (p\bar{c} + 2npM))^2}{t_1^2 + N_2 t_2^2}\right)\right), \tag{3.22}$$

avec

$$p \leq \left(\frac{nn_{\bar{B}}\rho^d\mathbb{V}_m}{t_1}\right)^m + N_2\left(\frac{n_{\bar{B}}n_B\mathbb{V}_m\Upsilon(d)}{t_2}\right)^m \text{ et } \bar{c} = t_1 + N_2 t_2,$$

et

$$\Upsilon(d) \text{ tel que } \Upsilon(d) \leq \upsilon = \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa\left(\frac{\kappa - 1}{\ln(\rho^{-1})e}\right)^{\kappa-1} \underset{\kappa\to\infty}{\sim} \frac{\kappa!}{\ln(\rho^{-1})^\kappa}.$$

Cette inégalité est directement issue de l'application d'une inégalité de type McDiarmid. Ce type d'inégalité peut s'appliquer si la variable que l'on cherche à contrôler est une fonction d'un nombre fini de variables aléatoires indépendantes, ce qui est bien le cas ici. En effet $\tilde{S}_{\mathscr{I}}^{[d]}$ s'exprime comme une fonction d'un nombre fini de variables approchées $\tilde{X}_t^{[d]}$, et chacune de ces variable $\tilde{X}_t^{[d]}$ s'exprime elle même comme une fonction d'un nombre fini d'innovations ε_t et d'une certaine variable $\bar{X}$ (cf équation)

On va donc présenter deux éléments préalables :

1. Les éléments de dénombrement nécessaire pour compter le nombre de variables aléatoires qui apparaissent dans $\tilde{S}_{\mathscr{I}}^{[d]}$.
2. L'inégalité de type McDiarmid que nous allons utiliser.

Dénombrement de variables aléatoires

La suite de cette section va nécessiter de compter les variables aléatoires qui apparaissent dans l'expression de $\tilde{S}_{\mathscr{I}}^{[d]}$. Dans un premier cas, on comptera le nombre de variables approchées (les variables notés $\tilde{X}_t^{[d]}$), dans un second cas les innovations (les variables ε_t). A cette fin, on introduit les cardinaux suivants :

— $n = \text{Card}(\mathscr{I})$, le nombre de variables aléatoires X_t apparaissant dans $S_{\mathscr{I}}$ (voir Section 3.2.3).

— $n_B = \text{Card}(\mathscr{B} \bigcup \{0\}) = \text{Card}(\mathscr{B}) + 1$,

— $n_{\bar{B}} = \text{Card}(\bar{\mathscr{B}})$

— $n_d = \text{Card}(\mathscr{V}(d\delta, t))$ le nombre d'innovations ε_t $\boldsymbol{\varepsilon}$ apparaissant dans l'approximation $\tilde{X}_t^{[d]}$ (voir Équation 3.4.2).

— $N_1 = \text{Card}(\bigcup_{t\in\mathscr{I}}\left(\mathscr{V}(\bar{\delta}, t)\right)) = \text{Card}(\bigcup_{t\in\mathscr{I}}\left(\bar{\mathscr{B}} + t\right))$ le nombre de variables approchées $\tilde{X}_t^{[d]}$ apparaissant dans $\tilde{S}_{\mathscr{I}}^{[d]}$.

— $N_2 = \text{Card}(\bigcup_{t\in\mathscr{I}}\left(\bigcup_{s\in\bar{\mathscr{B}}+t}\left(\mathscr{V}(d\bar{\delta}, t)\right)\right))$ le nombre d'innovations ε_t apparaissant dans $\tilde{S}_{\mathscr{I}}^{[d]}$.

On peut établir de manière immédiate

Lemme 3.3.

— $n \leq N_1 \leq nn_{\bar{B}}$ *et* $N_1 \leq N_2 \leq N_1 n_d$.

— $n_d = Card(\mathcal{V}(d\delta, t)) = \prod_{i=1}^{\kappa}(2d\delta_i + 1) \leq d^{\kappa}\prod_{i=1}^{\kappa}\left(2\delta_i + \frac{1}{d}\right) \leq d^{\kappa} n_B$.

Des bornes plus élevées pour N_1 et N_2 sont atteintes lorsque les unions ci-dessus impliquent des ensembles disjoints deux à deux. Néanmoins, ce n'est pas souvent le cas dans la pratique. Par exemple, dans les paramètres d'apprentissage automatique, les ensembles d'entraînement et de validation sont généralement des espaces connectés. Par conséquent, ajouter plus d'hypothèses sur la topologie de $\mathscr{I}$ devrait améliorer ces limites.

Une extension de l'inégalité de McDiarmid

Afin d'obtenir une inégalité de concentration pour $\tilde{S}_{\mathscr{I}}^{[d]}$, on a besoin d'une inégalité de McDiarmid spécifique. En effet, lorsque les hypothèses ($\mathbf{H_1^m}$) et ($\mathbf{H_2^m}$) sont vérifiées sans contraction absolue, l'hypothèse uniforme liée à la différence (telle que définie dans [Kut02]) n'est pas remplie, donc l'inégalité classique de McDiarmid [M+89] ne tient pas.
C'est pourquoi, on a besoin d'une version étendue de l'inégalité de McDiarmid qui tient même lorsque l'hypothèse de la différence bornée n'est vérifiée qu'avec une probabilité élevée. Plusieurs résultats de ce type existent ([Kut02; Com15; Kon14; War16]). Ici, on a choisi d'utiliser l'inégalité de McDiarmid étendue de [Com15].
On présente ci-dessous l'hypothèse "A-difference borned" qui correspond à l'hypothèse 1.2. de [Com15] et l'extension de l'inégalité de McDiarmid (théorème 2.1. de[Com15]).

Définition 3.9 (A-difference-bound). *$f : \prod_{i=1}^{N}\Omega_i \to \mathbb{R}$ est dit "A-difference bounded" par $(c_j)_{j\in[1,N]}$, s'il existe un ensemble $A \subset \prod_{i=1}^{N}\Omega_i$ tel que : pour $(w, w') \in A^2$ tels que w et w' diffèrent seulement sur la j-ème coordonnée , $|f(w) - f(w')| \leq c_j$.*

Lemme 3.4 (A-difference-bound McDiarmid, Théorème 2.1. issue [Com15]). *Si f est A-difference bounded par $(c_j)_{j\in[1,N]}$.*

$$\forall t > 0, \mathbb{P}(|f(X_1,\dots,X_n) - \mathbb{E}\left[f(X_1,\dots,X_n)|A\right]| \geq t) \leq 2\left(p + \exp\left(\frac{-2(t-p\bar{c})^2}{\sum_{j=1}^{N} c_j^2}\right)\right),$$

avec $\bar{c} = \sum_{j=1}^{N} c_i$ et $p = 1 - \mathbb{P}(A)$.

Borne de différence pour $\tilde{S}_{\mathscr{I}}^{[d]}$

Pour vérifier ces hypothèses (borne de différence forte ou "A-difference bound"), on doit borner la différence entre la statistique $S_{\mathscr{I}}$ et la statistique $S_{\mathscr{I}}$ lorsque l'une des variables aléatoires qui la composent est remplacée par une copie i.i.d. Pour traiter un tel cas, on doit introduire d'autres notations.

Nous rappelons que

$$\tilde{S}_{\mathscr{I}}^{[d]} = \sum_{t\in\mathscr{I}} \Phi((\tilde{X}_{t+s}^{[d]})_{s\in\bar{\mathscr{B}}}) = \sum_{t\in\mathscr{I}} \Phi(\tilde{H}^{[d]}(\bar{X}, (\varepsilon_s)_{s\in\mathcal{V}(d\delta,t)})).$$

Deux types de variables aléatoires sont impliqués dans $\tilde{S}_{\mathscr{I}}^{[d]}$.

— Les innovations ε_s (for $s \in \bigcup_{t\in\mathscr{I}} \mathcal{V}(d\bar{\delta}, t)$).

— La variable de remplissage $\bar{X}$.

Nous introduisons les notations suivantes.

— Pour tout t dans $\mathbb{Z}^{\kappa}$, $\tilde{X}_t^{[d]\prime} = \tilde{H}^{[d]}(\bar{X}', (\varepsilon_s)_{s\in\mathcal{V}(d\delta,t)})$ où $\bar{X}'$ suit la distribution μ_X et est indépendant de $(X_t)_{t\in\mathbb{Z}^{\kappa}}$, $\boldsymbol{\varepsilon}$ et $\bar{X}$.

— Pour tout t dans $\mathbb{Z}^{\kappa}$ et pour tout i dans $\mathcal{V}(d\bar{\delta}, t)$, $\tilde{X}_t^{[d,i]} = \tilde{H}^{[d]}(\bar{X}, (\varepsilon_s)_{s\in\mathcal{V}(d\delta,t)\setminus\{i\}} \bigcup \varepsilon_i')$ où ε_i' suit la distribution μ_ε et est indépendant de $(X_t)_{t\in\mathbb{Z}^{\kappa}}$, $\boldsymbol{\varepsilon}$ et $\bar{X}$.
Pour tout i dans $\mathbb{Z}^{\kappa}$ mais pas dans $\mathcal{V}(d\bar{\delta}, t)$, on pose $\tilde{X}_t^{[d,i]} = \tilde{X}_t^{[d]}$.

— Pour tout t dans $\mathbb{Z}^\kappa$, $\tilde{S}_{\mathscr{I}}^{[d]\prime} = \sum_{t\in\mathscr{I}} \Phi((\tilde{X}_{t+s}^{[d]\prime})_{s\in\tilde{\mathscr{B}}})$.

— Pour tout t dans $\mathbb{Z}^\kappa$ et pour tout i dans $\bigcup_{t\in\mathscr{I}} \mathcal{V}(d\bar{\delta},t)$, $\tilde{S_{\mathscr{I}}}^{[d,i]} = \sum_{t\in\mathscr{I}} \Phi(\tilde{X}_t^{[d,i]})$.

Lemme 3.5. *Soit $m\in\mathbb{N}\cup\{\infty\}$. Si $(\mathbf{H_1^m})$ et $(\mathbf{H_2^m})$ sont vérifiés,*

— *Pour tout t dans $\mathbb{Z}^\kappa$, $\|\tilde{X}_t^{[d]} - \tilde{X}_t^{[d]\prime}\|_m \le \rho^d \mathbb{V}_m$.*

— *Pour tout t et i dans $\mathbb{Z}^\kappa$.*

  — *Si $i\notin\mathcal{V}(d\bar{\delta},t)$, $\tilde{X}_t^{[d]} = \tilde{X}_t^{[d,i]}$.*

  — *Sinon, il y a un unique $c\in[0,d]$ tel que $i\in\mathcal{V}(c\delta,t)$ et $i\notin\mathcal{V}((c-1)d\delta,t)$ et qui vérifie :*

$$\|\tilde{X}_t^{[d]} - \tilde{X}_t^{[d,i]}\|_m \le \rho^c \mathbb{V}_m$$

.

Démonstration. On suppose $(\mathbf{H_1^m})$ et $(\mathbf{H_2^m})$.

— La démonstration du premier point est la même que le lemme 3.2.

— Premièrement, si $i\notin\mathcal{V}(d\delta,t)$, par définition de $\tilde{X}_t^{[d,i]}$, $\tilde{X}_t^{[d,i]} = Xttd$.

Deuxièmement, si $i\in\mathcal{V}(d\delta,t)$, on peut écrire $\mathcal{V}(d\delta,t) = \bigcup_{c=0}^{d} \mathcal{V}(c\delta,t)\backslash\mathcal{V}((c-1)\delta,t)$ avec $\mathcal{V}(0,t) = \{t\}$ et $\mathcal{V}(-1,t)=\emptyset$.
De plus, pour tout $c_1\neq c_2$, $\mathcal{V}(c_1\delta,t)\backslash\mathcal{V}((c_1-1)\delta,t)\bigcap\mathcal{V}(c_2\delta,t)\backslash\mathcal{V}((c_2-1)\delta,t)=\emptyset$.
En conséquence, il existe un unique c dans $[0,d]$ tel que $i\in\mathcal{V}(c\delta,t)$ et $i\notin\mathcal{V}((c-1)\delta,t)$. Alors avec le même argument que le lemme 3.2, on peut montrer que $\|\tilde{X}_t^{[d]} - \tilde{X}_t^{[d,i]}\|_m \le \rho^c\mathbb{V}_m$.

□

Nous pouvons maintenant borner la différence entre $\tilde{S}_{\mathscr{I}}^{[d]}$ et $\tilde{S}_{\mathscr{I}}^{[d]\prime}$ et $\tilde{S}_{\mathscr{I}}^{[d]}$ et $\tilde{S_{\mathscr{I}}}^{[d,i]}$.

Lemme 3.6 (Borne de différence pour $\tilde{S}_{\mathscr{I}}^{[d]}$)**.** *Soit $m\in\mathbb{N}\cup\{\infty\}$. Si $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$ et $(\mathbf{H_3})$ sont vérifiés, alors*

$$\|\tilde{S}_{\mathscr{I}}^{[d]} - \tilde{S}_{\mathscr{I}}^{[d]\prime}\|_m \le n n_{\bar{\mathrm{B}}} \rho^d \mathbb{V}_m.$$

et,

$$\|\tilde{S}_{\mathscr{I}}^{[d]} - \tilde{S_{\mathscr{I}}}^{[d,i]}\|_m \le n_{\bar{\mathrm{B}}}\mathbb{V}_m\left(1 + n_{\mathrm{B}}\kappa\sum_{c=1}^{d} c^{\kappa-1}\rho^c\right).$$

La preuve de ce lemme se trouve en annexe B.3.

Dans le lemme précédent, la quantité $\sum_{c=1}^{d} c^{\kappa-1}\rho^c$ apparaît, on montre dans le lemme suivant que nous pouvons borner cette quantité indépendamment de d.

Lemme 3.7. *Si $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$ et $(\mathbf{H_3})$ sont vérifiés, alors pour tout $d\in\mathbb{N}$, on a :*

$$\|\tilde{S_{\mathscr{I}}}^{[d,i]} - \tilde{S}_{\mathscr{I}}^{[d]}\|_m \le n_{\bar{\mathrm{B}}} n_{\mathrm{B}} \mathbb{V}_m \Upsilon(d).$$

La fonction Υ est définie par

$$\Upsilon(d) = \begin{cases} \kappa\left(\dfrac{1}{n_{\mathrm{B}}\kappa} + \dfrac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(1-\rho^{d+1}\displaystyle\sum_{i=0}^{\kappa-1}\frac{\left((d+1)\ln(\rho^{-1})\right)^i}{i!}\right)\right), & \text{if } d < \left\lfloor\dfrac{\kappa-1}{\ln(\rho^{-1})}\right\rfloor. \\ \kappa\left(\dfrac{1}{n_{\mathrm{B}}\kappa} + \dfrac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(1-e^{-(\kappa-1)}\displaystyle\sum_{i=0}^{\kappa-1}\frac{\left(\left\lfloor\frac{\kappa-1}{\ln(\rho^{-1})}\right\rfloor\ln(\rho^{-1})\right)^i}{i!}\right) + \left(\dfrac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}\right), & \text{if } d = \left\lfloor\dfrac{\kappa-1}{\ln(\rho^{-1})}\right\rfloor. \\ \kappa\left(\dfrac{1}{n_{\mathrm{B}}\kappa} + \dfrac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(1-\rho^{d}\displaystyle\sum_{i=0}^{\kappa-1}\frac{\left(d\ln(\rho^{-1})\right)^i}{i!}\right) + \left(\dfrac{\kappa-1}{\ln(\rho-1)e}\right)^{\kappa-1}\right), & \text{if } d > \left\lfloor\dfrac{\kappa-1}{\ln(\rho^{-1})}\right\rfloor. \end{cases}$$

avec

$$\Upsilon(d) \leq \upsilon = \frac{1}{n_{\mathrm{B}}} + \frac{\kappa!}{\ln(\rho^{-1})^{\kappa}} + \kappa\left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}.$$

On souligne que υ *est indépendant de* d *et* $\upsilon \underset{\kappa\to\infty}{\sim} \frac{\kappa!}{\ln(\rho^{-1})^{\kappa}}$ *(en utilisant la formule de Stirling).*

La preuve de ce lemme se trouve dans l'annexe B.4.

Inégalité de McDiarmid

On peut maintenant utiliser l'inégalité de McDiarmid pour $\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]}$ et prouver le théorème 3.4.

Ici, on fait la différence entre deux cas. Ou bien, on est dans le cadre d'une hypothèse de contraction uniforme. $(\mathbf{H_1^\infty})$ et $(\mathbf{H_2^\infty})$ sont vérifiés, et on peut utiliser une version classique de l'inégalité de McDiarmid. Le résultat est présenté dans le lemme 3.8.

Ou bien, on est dans le cadre d'une hypothèse de contraction faibles. $(\mathbf{H_1^m})$ et $(\mathbf{H_2^m})$ ne sont vérifiés que pour une valeur finie de m. Dans ce cas, l'hypothèse de différence bornée standard n'est pas remplie, on utilisera donc la notion de "A-difference bound" et on appliquera le Lemme 3.4. Le résultat est présenté dans le lemme 3.9.

Lemme 3.8. *Si* $(\mathbf{H_1^\infty})$,$(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$, *et pour* $d \in \mathbb{N}$, *alors on a :*

$$\begin{aligned}\mathbb{P}\left(|\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]}\right]| \geq \varepsilon\right) &\leq 2\exp\left(-2\frac{\varepsilon^2}{(n n_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty)^2 + \sum_{i=1}^{\mathrm{N}_2}\left(n_{\bar{\mathrm{B}}} n_{\mathrm{B}}\mathbb{V}_\infty\Upsilon(d)\right)^2}\right)\\ &\leq 2\exp\left(\frac{-2\varepsilon^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2\left(n^2\rho^{2d} + \mathrm{N}_2\left(n_{\mathrm{B}}\Upsilon(d)\right)^2\right)}\right).\end{aligned}$$

Démonstration. En utilisant le lemme 3.6 et 3.7 avec les hypothèses $(\mathbf{H_1^\infty})$ et $(\mathbf{H_2^\infty})$, on obtient

— $|\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]} - \tilde{\mathrm{S}}_{\mathscr{J}}^{[d]\prime}| \leq c_i$ with $c_i = n n_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty$ presque sûrement.

— $\forall i \in \bigcup_{t\in\mathscr{J}} \mathscr{V}(d\bar{\delta}, t), |\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]} - \tilde{\mathrm{S}}_{\mathscr{J}}^{[d,i]}| \leq c_i$ with $c_i = n_{\bar{\mathrm{B}}} n_{\mathrm{B}}\mathbb{V}_\infty\Upsilon(d)$. presque sûrement.

L'application de l'inégalité de McDiarmid [M+89] donne le résultat. □

Lemme 3.9. *Soit* $m \in \mathbb{N}$. *Si l'on suppose que* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et* $(\mathbf{H_4})$ *sont vérifiées, alors pour* $d \in \mathbb{N}$, *et* $t_1, t_2 > 0$, *il existe* $p \in (0,1)$, *tel que pour* $\varepsilon \geq 2npM + p\bar{c}$.

$$\mathbb{P}\left(|\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]}\right]| \geq \varepsilon\right) \leq 2\left(p + \exp\left(\frac{-2(\varepsilon - (p\bar{c} + 2npM))^2}{t_1^2 + \mathrm{N}_2 t_2^2}\right)\right),$$

avec

$$p \leq \left(\frac{n n_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_m}{t_1}\right)^m + \mathrm{N}_2\left(\frac{n_{\bar{\mathrm{B}}} n_{\mathrm{B}}\mathbb{V}_m\Upsilon(d)}{t_2}\right)^m \text{ et } \bar{c} = t_1 + \mathrm{N}_2 t_2.$$

La preuve de ce résultat est dans l'annexe B.5.

3.5 Inégalité de concentration pour $\mathrm{S}_{\mathscr{J}}$

3.5.1 Inégalité de concentration pour $\mathrm{S}_{\mathscr{J}}$

L'objectif de cette section est d'établir une inégalités de concentration pour $\mathrm{S}_{\mathscr{J}}$. Il peut être réalisé en utilisant les inégalités de concentration de la section précédente (impliquant $\tilde{\mathrm{S}}_{\mathscr{J}}^{[d]}$) et le lemme 3.2.

Lemme 3.10. INÉGALITÉ DE CONCENTRATION POUR $S_{\mathscr{J}}$, CONTRACTION UNIFORME *Soit $d \in \mathbb{N}$. Si $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ et $(\mathbf{H_3})$ sont vérifiés, alors on a la relation suivante pour $\varepsilon > 2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty$.*

$$\mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\left(\varepsilon - 2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2\left(n^2\rho^{2d} + \mathrm{N}_2\left(n_{\mathrm{B}}\Upsilon(d)\right)^2\right)}\right). \tag{3.23}$$

Soit $m \in \mathbb{N}$. Si $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ et $(\mathbf{H_4})$ sont vérifiés, alors pour, $t_1, t_2 > 0$, il existe $p \in (0,1)$ tel que pour $\varepsilon > 4npM + 2p\bar{c}$.

$$\mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq \varepsilon\right) \leq 2\left(p + \exp\left(\frac{-2(\frac{\varepsilon}{2} - (p\bar{c} + 2npM))^2}{t_1^2 + \mathrm{N}_2 t_2^2}\right)\right) + \left(\frac{2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_m}{\varepsilon}\right)^m, \tag{3.24}$$

avec

$$p \leq \left(\frac{nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_m}{t_1}\right)^m + \mathrm{N}_2\left(\frac{n_{\bar{\mathrm{B}}}n_{\mathrm{B}}\mathbb{V}_m\Upsilon(d)}{t_2}\right)^m \text{ and } \bar{c} = t_1 + \mathrm{N}_2 t_2.$$

Démonstration.

Si $(\mathbf{H_1^\infty})$,$(\mathbf{H_2^\infty})$ et $(\mathbf{H_3})$ sont vérifiés. En utilisant le corollaire 3.6, on a presque sûrement :

$$\begin{aligned}
|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| &= |S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]} + \tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right] + \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right] - \mathbb{E}[S_{\mathscr{J}}]| \\
&\leq |S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}| + |\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathscr{J}}^{[d]}]| + |\mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right] - \mathbb{E}[S_{\mathscr{J}}]| \\
&\leq |\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathscr{J}}^{[d]}]| + 2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty.
\end{aligned}$$

Par conséquence, pour $\varepsilon > 2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty$.

$$\begin{aligned}
\mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq \varepsilon\right) &\leq \mathbb{P}\left(|\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right]| \geq \varepsilon - 2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty\right) \\
&\leq 2\exp\left(\frac{-2\left(\varepsilon - 2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_\infty\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2\left(n^2\rho^{2d} + \mathrm{N}_2\left(n_{\mathrm{B}}\Upsilon(d)\right)^2\right)}\right).
\end{aligned}$$

Si $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$ et $(\mathbf{H_3})$ sont vérifiés. En utilisant le corollaire 3.6 et l'inégalité de Markov,

$$\forall t > 0, \mathbb{P}\left(|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}| \geq t\right) \leq \left(\frac{nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_m}{t}\right)^m.$$

Alors

$$\begin{aligned}
\mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq \varepsilon\right) = \mathbb{P}\left(|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]} + \tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}[S_{\mathscr{J}}]| \geq \varepsilon\right) &\leq \mathbb{P}\left(|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}| + |\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}[S_{\mathscr{J}}]| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(|S_{\mathscr{J}} - \tilde{S}_{\mathscr{J}}^{[d]}| \geq \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}[S_{\mathscr{J}}]| \geq \frac{\varepsilon}{2}\right) \\
&\leq 2\left(p + \exp\left(\frac{-2(\frac{\varepsilon}{2} - (p\bar{c} + 2npM))^2}{t_1^2 + \mathrm{N}_2 t_2^2}\right)\right) + \left(\frac{2nn_{\bar{\mathrm{B}}}\rho^d\mathbb{V}_m}{\varepsilon}\right)^m.
\end{aligned}$$

□

Remarque 3.10. *Le choix du paramètre d est un compromis entre la qualité de l'approximation et le nombre de variables aléatoires impliquées dans l'inégalité de McDiarmid. En effet, lorsque d augmente, l'erreur d'approximation diminue (voir le lemme 3.6), mais le nombre de variables aléatoires N_2 augmente (voir la borne pour N_2 dans le lemme 3.3).*

3.5.2 Optimisation du paramètre d

Les équations (3.23) et (3.24) dans le lemme 3.10 ont été établies pour tout $d \in \mathbb{N}$ et toute valeur positive de t_1, t_2. Par conséquent, on peut choisir une valeur appropriée pour chacun de ces paramètres afin d'améliorer nos bornes.
De plus, la quantité N_2 n'est pas facile à interpréter et encore plus à estimer. Ainsi, dans les théorèmes suivants, on fixe les paramètres d, t_1, t_2 et on remplace N_2 par une majoration correspondant au pire des cas.

Théorème 3.5. INÉGALITÉ DE CONCENTRATION AMÉLIORÉE POUR $S_{\mathscr{I}}$, CONTRACTION UNIFORME *Si* $(\mathbf{H_1^\infty})$,$(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés, alors* $\varepsilon > 2n_{\bar{B}}\mathbb{V}_\infty$ *:*

$$\mathbb{P}\left(|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\left(\varepsilon - 2n_{\bar{B}}\mathbb{V}_\infty\right)^2}{\left(n_{\bar{B}}\mathbb{V}_\infty\right)^2\left(1 + n_{\bar{B}}n_B^3\Upsilon(d)^2 n d^\kappa\right)}\right),$$

avec $K_\rho(\infty) = \frac{1}{\ln(\rho^{-1})}$, $\tilde{d} = K_\rho(\infty)\ln(n)$ *and* $d = \lceil \tilde{d} \rceil = \left\lceil \frac{\ln(n)}{\ln(\rho^{-1})} \right\rceil$.

Démonstration. On utilise le lemme précédent 3.10 et on pose $\tilde{d} = \frac{\ln(n)}{\ln(\rho^{-1})}$ et $d = \lceil \tilde{d} \rceil$.
Par conséquent

$$d^\kappa = \left\lceil \frac{\ln(n)}{\ln(\rho^{-1})} \right\rceil^\kappa.$$
$$\rho^d \leq \rho^{\tilde{d}} = \exp(-\ln(\rho^{-1})d) = \frac{1}{n}.$$

En partant du Lemme 3.3, on a $N_2 \leq N_1 n_d \leq n_{\bar{B}} n_B d^\kappa n$. Donc, selon le Lemme 3.10, on a :

$$\forall \varepsilon > 2nn_{\bar{B}}\rho^d\mathbb{V}_\infty, \mathbb{P}\left(|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\left(\varepsilon - 2nn_{\bar{B}}\rho^d\mathbb{V}_\infty\right)^2}{\left(n_{\bar{B}}\mathbb{V}_\infty\right)^2\left(n^2\rho^{2d} + N_2\left(n_B\Upsilon(d)\right)^2\right)}\right)$$
$$\Rightarrow \forall \varepsilon > 2n_{\bar{B}}\mathbb{V}_\infty, \mathbb{P}\left(|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\left(\varepsilon - 2n_{\bar{B}}\mathbb{V}_\infty\right)^2}{\left(n_{\bar{B}}\mathbb{V}_\infty\right)^2\left(1 + n_{\bar{B}}n_B^3\Upsilon(d)^2 n d^\kappa\right)}\right).$$

□

Théorème 3.6. INÉGALITÉ DE CONCENTRATION AMÉLIORÉE POUR $S_{\mathscr{I}}$, CONTRACTION FAIBLE *Si* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et*$(\mathbf{H_4})$ *sont vérifiés, alors pour* $\varepsilon \geq 4n_{\bar{B}}n_B d^{2\kappa}\rho^{-1}n^{\frac{2}{m}}L(n)$

$$\mathbb{P}\left(|S_{\mathscr{I}} - \mathbb{E}[S_{\mathscr{I}}]| \geq \varepsilon\right)$$
$$\leq 2\exp\left(\frac{-2\rho^2\left(\frac{\varepsilon}{2} - 2n_{\bar{B}}n_B d^{2\kappa}\rho^{-1}n^{\frac{2}{m}}L(n)\right)^2}{\left(n_{\bar{B}}\mathbb{V}_m n^{\frac{2}{m}}\right)^2\left(1 + nn_{\bar{B}}n_B^3\Upsilon(d)^2 d^\kappa\right)}\right) + \frac{\rho^m}{n}\left(2n_{\bar{B}}n_B d^\kappa + \left(\frac{\mathbb{V}_m}{2n_B d^{2\kappa}L(n)}\right)^m\right).$$

Avec $K_\rho(m) = \frac{1-\frac{1}{m}}{\ln(\rho^{-1})}$, $\tilde{d} = K_\rho(m)\ln(n)$, $d = \lceil \tilde{d} \rceil = \left\lceil \frac{(1-\frac{1}{m})\ln(n)}{\ln(\rho^{-1})} \right\rceil$.
et $L(n) = \left(\frac{n_{\bar{B}}\mathbb{V}_m}{\rho}\right)\left(\frac{1}{d^\kappa(n_{\bar{B}}n_B d^\kappa)^{\frac{1}{m}}n} + n_{\bar{B}}n_B^2\Upsilon(d)\right) + \frac{2M}{d^\kappa n^{\frac{2}{m}}}$.

La preuve se trouve dans l'annexe B.6.

Remarque 3.11. *Dans les théorèmes 3.5 et 3.6, les quantités* $n_{\bar{B}}, n_{\mathscr{B}}, \mathbb{V}_m, \mathbb{V}_\infty, \kappa$ *et* ρ *sont des constantes et la fonction* $\Upsilon(d)$ *peut toujours être contrôlée par une constante* υ *indépendamment de* d *(voir lemme 3.7).*

Remarque 3.12. *Dans le théorème 3.6, un terme additif supplémentaire apparaît. Ce terme décroît rapidement avec* n *et n'est donc souvent pas préjudiciable en pratique.*
De plus, l'ajout d'hypothèses appropriées sur m *peut conduire à une inégalité de concentration entièrement exponentielle.*

3.5.3 Limites de déviation attendues

Les limites de déviation attendues peuvent être obtenues à partir des théorèmes précédents. Dans ce cas, on veut borner $\mathbb{E}[|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]|]$.

Corollaire 3.7. *Si on suppose que* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés, on a :*

$$\mathbb{E}[|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]|] \leq n_{\bar{\mathrm{B}}} \mathbb{V}_\infty \left(2 + \sqrt{\frac{\pi}{2}\left(1 + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^\kappa n\right)}\right),$$

avec $K_\rho(\infty) = \frac{1}{\ln(\rho^{-1})}$ *et* $d = \lceil K_\rho(\infty)\ln(n)\rceil$.

Démonstration.

$$\begin{aligned}
\mathbb{E}[|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]|] &= \int_0^\infty \mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq t\right) dt \\
&= \int_0^{2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty} \mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq t\right) dt + \int_{2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty}^\infty \mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq t\right) dt \\
&\leq 2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty + \int_{2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty}^\infty 2\exp\left(\frac{-2\left(t - 2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2\left(1 + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^\kappa n\right)}\right) dt \\
&\leq 2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty + n_{\bar{\mathrm{B}}}\mathbb{V}_\infty \sqrt{\frac{\pi}{2}\left(1 + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^\kappa n\right)}.
\end{aligned}$$

□

Corollaire 3.8. *Si* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et*$(\mathbf{H_4})$ *sont vérifiés, alors*

$$\begin{aligned}
\mathbb{E}[|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]|] \leq & 2\frac{n_{\bar{\mathrm{B}}}\mathbb{V}_m n^{\frac{2}{m}}}{\rho}\sqrt{\frac{\pi}{2}\left(1 + n n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^\kappa\right)} + 4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) \\
& + 2\rho^m M\left(2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^\kappa + \left(\frac{\mathbb{V}_m}{2n_{\mathrm{B}} d^{2\kappa} L(n)}\right)^m\right),
\end{aligned}$$

avec $K_\rho(m) = \frac{1-\frac{1}{m}}{\ln(\rho^{-1})}$, $d = \lceil K_\rho(m)\ln(n)\rceil$ *et* $L(n) = \left(\frac{n_{\bar{\mathrm{B}}}\mathbb{V}_m}{\rho}\right)\left(\frac{1}{d^\kappa (n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^\kappa)^{\frac{1}{m}} n} + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^2 \Upsilon(d)\right) + \frac{2M}{d^\kappa n^{\frac{2}{m}}}$.

Démonstration.

$$\begin{aligned}
&\mathbb{E}[|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]|] \\
&= \int_0^\infty \mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq t\right) dt \\
&= \int_0^{4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)} \mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq t\right) dt + \int_{4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}^\infty \mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq t\right) dt \\
&\leq 4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) + \int_{4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}^{2nM} \mathbb{P}\left(|S_{\mathscr{J}} - \mathbb{E}[S_{\mathscr{J}}]| \geq t\right) dt \\
&(\text{car } S_{\mathscr{J}} \leq nM \text{ (hypothèse } (\mathbf{H_4})\text{)}) \\
&\leq 4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) + \int_{4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}^{2nM} 2\exp\left(\frac{-2\rho^2\left(\frac{t}{2} - 2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_m n^{\frac{2}{m}}\right)^2\left(1 + n n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^\kappa\right)}\right) dt \\
&+ 2M\rho^m\left(2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^\kappa + \left(\frac{\mathbb{V}_m}{2n_{\mathrm{B}} d^{2\kappa} L(n)}\right)^m\right).
\end{aligned}$$

De plus,

$$\int_{4n_{\bar{\mathrm{B}}}n_{\mathrm{B}}d^{2\kappa}\rho^{-1}n^{\frac{2}{m}}\mathrm{L}(n)}^{2n\mathrm{M}} 2\exp\left(\frac{-2\rho^2\left(\frac{t}{2}-2n_{\bar{\mathrm{B}}}n_{\mathrm{B}}d^{2\kappa}\rho^{-1}n^{\frac{2}{m}}\mathrm{L}(n)\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_m n^{\frac{2}{m}}\right)^2\left(1+nn_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\Upsilon(d)^2d^{\kappa}\right)}\right)dt$$
$$=4\int_0^{2n\mathrm{M}}\exp\left(\frac{-2\rho^2t^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_m n^{\frac{2}{m}}\right)^2\left(1+nn_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\Upsilon(d)^2d^{\kappa}\right)}\right)dt \leq 2\frac{n_{\bar{\mathrm{B}}}\mathbb{V}_m n^{\frac{2}{m}}}{\rho}\sqrt{\frac{\pi}{2}\left(1+nn_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\Upsilon(d)^2d^{\kappa}\right)}.$$

C'est pourquoi,

$$\mathbb{E}[|\mathrm{S}_{\mathscr{J}}-\mathbb{E}[\mathrm{S}_{\mathscr{J}}]|] \leq 4n_{\bar{\mathrm{B}}}n_{\mathrm{B}}d^{2\kappa}\rho^{-1}n^{\frac{2}{m}}\mathrm{L}(n)+2\frac{n_{\bar{\mathrm{B}}}\mathbb{V}_m n^{\frac{2}{m}}}{\rho}\sqrt{\frac{\pi}{2}\left(1+nn_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\Upsilon(d)^2d^{\kappa}\right)}$$
$$+2\rho^m\mathrm{M}\left(2n_{\bar{\mathrm{B}}}n_{\mathrm{B}}d^{\kappa}+\left(\frac{\mathbb{V}_m}{2n_{\mathrm{B}}d^{2\kappa}\mathrm{L}(n)}\right)^m\right).$$

□

En utilisant ces inégalités de moment, nous pouvons également analyser le comportement limite de $\mathbb{E}[|\mathrm{S}_{\mathscr{J}}-\mathbb{E}[\mathrm{S}_{\mathscr{J}}]|]$.

Corollaire 3.9.

— *Si* $(\mathbf{H_1^\infty})$, $(\mathbf{H_2^\infty})$ *et* $(\mathbf{H_3})$ *sont vérifiés, alors*

$$\mathbb{E}[|\mathrm{S}_{\mathscr{J}}-\mathbb{E}[\mathrm{S}_{\mathscr{J}}]|] \leq \mathrm{G}_1(\kappa,\rho,\mathbb{V}_\infty,n_{\bar{\mathrm{B}}},n_{\mathrm{B}},n) \underset{n\to\infty}{\sim} n_{\mathrm{B}}n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\upsilon\sqrt{\frac{\pi}{2}n_{\mathrm{B}}n_{\bar{\mathrm{B}}}(\mathrm{K}_\rho(\infty))^\kappa n\ln(n)^\kappa},$$
$$\underset{n\to\infty}{=}\mathscr{O}\left(\sqrt{n\ln(n)^\kappa}\right),$$

avec $\mathrm{K}_\rho(\infty)=\frac{1}{\ln(\rho^{-1})}$.

— *Si* $(\mathbf{H_1^m})$, $(\mathbf{H_2^m})$, $(\mathbf{H_3})$ *et*$(\mathbf{H_4})$ *sont vérifiés, alors*

$$\mathbb{E}[|\mathrm{S}_{\mathscr{J}}-\mathbb{E}[\mathrm{S}_{\mathscr{J}}]|] \leq \mathrm{G}_2(\kappa,\rho,\mathbb{V}_m,n_{\bar{\mathrm{B}}},n_{\mathrm{B}},n) \underset{n\to\infty}{\sim} \frac{n_{\bar{\mathrm{B}}}n_{\mathrm{B}}\mathbb{V}_m\upsilon n^{\frac{2}{m}}}{\rho}\sqrt{2\pi n_{\bar{\mathrm{B}}}n_{\mathrm{B}}\mathrm{K}_\rho(m)^\kappa\ln(n)^\kappa n}$$
$$\underset{n\to\infty}{=}\mathscr{O}\left(n^{\frac{2}{m}}\sqrt{n\ln(n)^\kappa}\right),$$

avec $\mathrm{K}_\rho(m)=\frac{1-\frac{1}{m}}{\ln(\rho^{-1})}$.

Démonstration. On utilise les corollaires 3.7 et 3.8 et la borne$\forall d\in\mathbb{N},\Upsilon(d)\leq\upsilon$ (voir Lemme 3.7). □

Remarque 3.13. *Le théorème 3.6, les corollaires 3.8 et 3.9 avec les hypothèses* $(\mathbf{H_1^m})$ *et* $(\mathbf{H_2^m})$ *ne sont utiles que si* $m>4$. *Si ce n'est pas le cas* $\lim_{n\to+\infty}n^{\frac{2}{m}-\frac{1}{2}}\neq 0$, *par conséquent* $lim_{n\to+\infty}\mathbb{E}\left[\frac{|\mathrm{S}_{\mathscr{J}}-\mathbb{E}[\mathrm{S}_{\mathscr{J}}]|}{n}\right]\neq 0$. *Dans le contexte de la théorie de l'apprentissage, cela signifie que le risque empirique peut ne pas converger vers le risque théorique (voir la section 3.3.2).*

Deuxième partie

Prévision de séries temporelles multiples et application au cadre de la prévision des ventes

Chapitre 4

Qu'est ce qu'une bonne prédiction des ventes ?

Dans ce chapitre, on détaillera les objectifs d'une prévision [1] de la demande, et on se demandera ce qui fait une bonne prévision. On présentera également ce qui rend une telle prévision compliquée. Le but n'est pas ici de refaire un cours exhaustif sur la prévision des ventes. De nombreux ouvrages existent sur ce sujet. On peut citer notamment [Van21] pour une revue des techniques statistique et Data Science utilisées pour la prédiction des ventes. Cependant, le cadre de Cdiscount, et celui du E-commerce diffère par endroit avec le cadre des prévisions des ventes classique. On s'intéressera donc à pointer dans ce chapitre les différences avec le cas usuel.

4.1 Pourquoi prédire les ventes ?

Les chaînes logistiques sont soumises aux variations de la demande. La demande pour un produit peut évoluer du fait de nombreux facteurs. C'est particulièrement le cas pour un site d'E-Commerce, qui propose de nombreux produits ayant une flexibilité-prix importante, et où les prix et de nombreux autres paramètres varient en permanence. Face à de telles variations, les logisticiens doivent prendre des marges de sécurité afin d'éviter les ruptures de stock. Ces marges de sécurités sont cependant également un problème pour les logisticiens, qui cherchent en général à limiter les stocks. En effet, le stock représente un coût important pour une entreprise, d'une part à cause des coûts du stockage en lui-même (entrepôt, assurance, etc...), d'autre part à cause du risque de non-vente. Ici, l'E-Commerce a un avantage : la plupart des produits ne sont pas périssables. Cependant, dans certains cas, il peut exister des obsolescences matérielles, certains produits peuvent se démoder. Aussi, dans ce contexte, les stocks invendus doivent souvent être bradés. Prédire la demande est donc nécessaire pour pouvoir réapprovisionner correctement les entrepôts. Une prévision plus fiable permet de réduire les marges de sécurité, et d'éviter également les ruptures. Notons cependant que la prévision de la demande n'est pas toujours nécessaire pour les différents produits. D'une part, parce que l'horizon de réapprovisionnement joue un rôle important : s'il est possible de réapprovisionner un produit en quelques jours, il n'est pas nécessaire d'avoir des prévisions de la demande sur plusieurs mois. D'autre part, si le produit est facilement remplaçable, une rupture de stock n'est pas nécessairement problématique. Par exemple, le consommateur est généralement indifférent au modèle ou à la marque de certains produits bricolages (piles, tournevis, etc...). Prédire les ventes de chacun de ces produits est donc moins important. On peut remarquer qu'il existe une différence entre la prévision des ventes et la prévision de la demande. Les ventes peuvent être impactées par la présence de stock ou non, notamment en cas de ruptures. La demande n'est au contraire pas affectée par l'état du stock. Lorsqu'il y a rupture de stock, il existe une demande non satisfaite. L'objectif pour une entreprise

1. On parlera indifféremment de prévision, ou de prédiction des ventes. Il ne semble pas qu'il y ait de termes dominants dans la littérature, ni de fortes différences entre les deux.

est de prédire correctement la demande, et non les ventes, pour réduire cette demande non satisfaite. Cependant, la demande n'est jamais complètement observée, et ne peut provenir que d'un modèle statistique.

4.2 Intérêt et limites du cadre mathématique

Le cadre mathématique classique sous lequel on présente généralement la prévision des ventes est celui décrit et utilisé en introduction. Ce cadre analogue au cadre de la prédiction de série temporelle dans les précédents chapitres. Considérons un ensemble de produits, numérotés de 1 à d, appelé le *catalogue* Si l'on note $(x_{i,t}) \in \mathbb{R}$ et h l'horizon de prédiction et que l'on suppose connues toutes les ventes jusqu'à l'horizon t_0, le but est de trouver une estimation $\hat{x}_{i,t+h}$ telle que

$$\hat{x}_{i,t_0+h} = \underset{x}{\operatorname{argmin}}\, \mathbb{E}[\mathrm{L}(x, x_{i,t_0+h}) | \forall j, \forall t < t_0, x_{j,t}] \tag{4.1}$$

.

L étant une perte de même nature que la perte considérée au chapitre 1 et suivants.

Cependant, cette façon de présenter les choses est un peu réductrice. En effet, le vrai problème de prévision des ventes concerne plusieurs semaines successives. Il faudrait plutôt considérer l'ensemble des ventes, jusqu'à l'horizon h, que l'on suppose en général être le temps nécessaire au réapprovisionnement. Toujours pour des ventes $(x_{i,t})$, on cherche à résoudre :

$$\hat{y}_{i,t_0+h} = \underset{y}{\operatorname{argmin}}\, \mathbb{E}[\mathrm{L}(y, \sum_{t=t_0+1}^{t_0+h} x_{i,t}) | \forall j, \forall t < t_0, x_{j,t}] \tag{4.2}$$

Ici, on a une approche globale, qui a l'avantage de demander de ne prédire que des valeurs agrégées sur plusieurs semaines successives, ce qui est relativement plus simple que de prédire chaque semaine indépendamment. L'approche globale 4.2 ne semble cependant que peu étudiée par rapport à l'approche point par point 4.1. On peut avancer plusieurs explications pour justifier ce phénomène. D'abord 4.1 est plus proche des problèmes classiques de séries temporelles. L'approche point par point a quelques avantages d'un point de vue également computationnel : elle ne présente pas d'effet de bord, permet d'utiliser un plus grand nombre de point de données. Enfin, d'un point de vue pratique, la prédiction semaine par semaine a également l'avantage d'être plus facile à interpréter, puisqu'une variation de prévision sera directement lié à un effet hebdomadaire, comme la présence d'un évènement particulier comme les soldes par exemple.

La question de l'optimisation réelle des commandes en fonction des prévisions des ventes est un problème de recherche opérationnelle assez complexe qui ne sera pas traité ici. [BM12] offre un exemple d'articulation entre prévision et optimisation des commandes. La notion clef utilisée est celle de stock du sûreté, c'est à dire la marge de sécurité prise en terme de stocks. Cependant, la façon de définir le stock de sûreté peut varier, tout comme l'objectif que l'on cherche à atteindre et la méthode pour l'évaluer. En particulier, il existe de multiples façons d'évaluer le sur-stockage et le niveau de service. [BCC21] présente une revue des méthodes pour définir ce stock, ainsi que des objectifs associés.

4.3 Métriques

Dans cette section, on s'intéressera à des exemples spécifiques de perte L, et on s'interrogera sur le rôle et l'intérêt de la notion de métrique dans le cadre de la prédiction des ventes.

Pour évaluer la qualité d'une prévision de la demande point par point, on peut faire appel à différentes métriques. Parmi ces métriques, les plus populaires sont le MAPE (Mean Absolute Percentage Error), le %MAE (Scaled Mean Absolute Error) et le RMSE(Root Mean Squared Error). Sur une période de temps [1, T] ,pour des prévisions $(\hat{x}_{i,t})$ et des ventes $(x_{i,t})$, on définit ces métriques de la façon suivante :

— MAPE = $\frac{1}{d}\sum_{i=1}^{d}\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\frac{|\hat{x}_{i,t}-x_{i,t}|}{x_{i,t}}$ Cette métrique, classique en apprentissage ne devrait pas être utilisée en général pour la prévision . En effet, elle accorde un poids démesuré aux erreurs lorsque les ventes sont faibles, ce qui est très souvent le cas pour la prévision des ventes.

— %MAE = $\frac{1}{d}\sum_{i=1}^{d}\frac{\sum_{t=1}^{\mathrm{T}}|\hat{x}_{i,t}-x_{i,t}|}{\sum_{t=1}^{\mathrm{T}}x_{i,t}}$ Cette métrique est très simple à utiliser, et sera appliquée pour l'évaluation des résultats dans la suite. Optimiser cette métrique, c'est optimiser la médiane. La médiane étant généralement inférieure à la moyenne, l'optimisation de cette métrique donne souvent des prédictions qui sous-estiment légèrement les ventes. Notons que cette métrique est parfois appelée MAPE lorsqu'elle est exprimée en pourcentage.

— RMSE = $\frac{1}{d}\sum_{i=1}^{d}\sqrt{\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}(\hat{x}_{i,t}-x_{i,t})^2}$. Cette métrique est également un grand classique. Optimiser cette métrique, c'est optimiser la moyenne. Cependant, la moyenne est sensible à la présence d'anomalies (outliers), qui sont souvent nombreuses dans les données.

On peut également pondérer ces métriques en y introduisant la notion de prix.

Ces trois métriques ont l'avantage d'être facilement interprétables, et peuvent être expliquées facilement à des responsables logistiques. Ces métriques "intuitives" sont utiles pour présenter les performances des modèles.

Cependant, il existe d'autres métriques intéressantes lors de la conception des modèles. Ces métriques sont moins facile à interpréter, mais tendent à donner de meilleurs résultats. On présentera au chapitre 3 un exemple de telle métrique avec la loss-Poisson. Un autre exemple souvent utilisé durant cette thèse est la loss %MAE déséquilibrée par un facteur $\gamma > 1$:

$$\%\mathrm{MAED}(\gamma) = \frac{1}{d}\sum_{i=1}^{d}\frac{\sum_{t=1}^{\mathrm{T}}|\hat{x}_{i,t}-x_{i,t}|(1+(\gamma-1)\mathbb{1}_{x_{i,t}>\hat{x}_{i,t}})}{\sum_{t=1}^{\mathrm{T}}x_{i,t}}$$

L'idée est de pénaliser plus fortement la sous-prédiction pour contrer la tendance naturelle des algorithmes de Machine Learning à sous-prédire les ventes. Ce type de métrique peut être utilisée comme perte L dans les algorithmes de machine learning, ou bien pour effectuer de la selection de modèle comme au chapitre 2.

4.4 Importance de l'interprétabilité

Les prévisions des ventes sont utilisées pour piloter des commandes de produits à différents fournisseurs. Bien que certaines entreprises passent une part de leurs commandes de manière automatique, en pratique, la commande est souvent ordonné par un humain. Chez CDiscount par exemple, un algorithme suggère des ordres de commandes à partir des prévision. L'humain est alors là pour valider ou refuser ces suggestions. Cependant, pour que cette validation puisse se faire, il est nécessaire que la personne en question puisse comprendre ce qui fait une prévision, quelles sont les variables utilisées, ce qui rentre dans le calcul, et ce qui n'y participe pas. On va donc chercher naturellement à rendre ces prévisions interprétables.

Cela rentre naturellement en conflit avec la complexité inhérente aux modèles de ventes les plus performants, dont le manque d'interprétabilité est notoire [CPC19]. Dans le cadre de la prévision des ventes, 3 causes majeures rendent l'interprétabilité des modèles complexe :

— Non Linéarité : Les modèles les plus performants sont non-linéaires. En effet, les effets observés ne sont en général pas linéaires. Par exemple, une promotion sur un produit de 20% n'aura pas un impact doublé par rapport à une promotion de 10%.

— Séparabilité : De nombreuses variables externes sont co-intégrées. Ainsi, durant des épisodes de soldes, on observe à la fois une augmentation de la fréquentation du site web, et des diminutions de prix. Il devient alors difficile de séparer les effets, et d'attribuer tant de ventes à un effet, et tant de ventes à un autre.

— Multiplicité des niveaux : L'importance d'une variable interprétative donnée dépend très fortement du produit dont on observe les ventes, et notamment de sa catégorie, son niveau

de ventes, sa marque. Ces paramètres catégoriels doivent être pris en compte par le modèle. On doit alors soit construire un modèle pour chaque catégorie, soit intégrer d'une manière ou d'une autre ces informations hiérarchiques au modèle. Dans les deux cas, cela rend l'interprétation des prévisions particulièrement complexe.

Cependant, la performance des modèles de type "black-box" rend ces derniers très utiles. Par ailleurs, les utilisateurs n'ont pas nécessairement besoin d'une pure interprétabilité mathématique, simplement d'une compréhension des phénomènes pris en compte dans l'élaboration d'une prédiction donnée. Cela pousse à deux choses : Créer des index permettant de mieux comprendre comme l'index de compétitivité du Chapitre 6 , et permettre à l'utilisateur de rejouer certaines prédictions avec d'autres paramètres et features.

4.5 Cadre industriel de la thèse et typologie des méthodes de prédictions

4.5.1 Volumétrie

Dans le cadre de cette thèse, on s'est intéressé aux ventes de différents produits de l'entreprise CDiscount. Ces ventes ont été agrégées par semaine et par produit. L'objectif était de réaliser une prévision de la demande à moyen et long terme, c'est-à-dire typiquement entre 4 et 52 semaines. Le catalogue total comprends en théorie plus de 2 100 000 de produits répartis sur une petite dizaine d'entrepot. En pratique, il n'y a en moyenne qu'autour de 110 000 produits actifs à un instant donné, chaque produit étant assigné au plus à 3 entrepôts. Par ailleurs, on peut encore réduire la taille du catalogue si on enlève les produits ayant les plus faibles rotations, c'est à dire des produits ayant peu de ventes par mois (moins d'une dizaine). En faisant ainsi, on obtient un catalogue 'noyau' comprenant les produits intéressants représentants environ 96% du volume d'affaire. Ce sont ces produits sur lesquels on va essayer de faire de la prévision des ventes.

En croisant différentes bases de données, on peut obtenir un historique de vente remontant à 2015, même si les données avant 2018 sont moins fiables, et ne rendent pas exactement compte de la même notion de ventes. Les autres données disponibles ont des périmètres d'historisation variables également. Globalement, les données pour les années 2019,2020 et 2021 sont assez complètes et fiables, ce qui n'est pas nécessairement le cas pour des données plus anciennes.

Tout cela a pour conséquence qu'on est dans un problème avec peu d'historique, et beaucoup de séries. Ce n'est pas à proprement parler un problème de grande dimension, mais il faut noter que des approches globales de type VAR 'naif' semblent peu adaptées pour cette raison.

4.5.2 Les prévisions des ventes chez CDiscount

Lors du démarrage des travaux de cette thèse, les ventes étaient prédites par des logiciels conçues par des entreprises spécialisées dans la gestion d'inventaire : Planipe et Relex. Ces entreprises se basent sur les ventes passées d'un produits, et agrègent différents prédicteurs classiques pour effectuer leurs prédictions. Parmi ces prédicteurs, on peut citer différents types de lissages exponentiels (simple, double et triple lissages exponentiels, saisonnalisés ou non), des méthodes auto-régressives et des méthodes additives. Il est apparu assez tôt qu'il était inutile de tenter d'améliorer les prédictions sur ce terrain, étant donné l'expertise de ces entreprises.

Cependant, 2 axes peuvent être exploités pour tenter d'améliorer les prévisions . Premièrement, on dispose d'informations internes à CDiscount qui permettaient de mieux expliquer les variations de ventes. On présentera ces différentes features en section 4.5.4.

Deuxièmement on souhaite exploiter une approche globale de prédiction des ventes, au lieu d'effectuer la prédiction séries par séries. A ce titre, différentes approches ont été essayés au cours de la thèse, même si elle ne seront pas toutes présentées dans le présent manuscrit.

4.5.3 Typologie des méthodes de prédiction

Essayons de dresser une typologie de différentes méthodes de prévision. On reprend les notations de la section 4.2, à savoir qu'on observe des ventes $(x_{i,t})$ dont on cherche à effectuer une prédiction $(\hat{x}_{i,t+h})$. On note également $\theta_{i,t}$ les variables externes concernant un produit i à la date t.

— L'approche auto-régressive : Le principe est relativement simple : On cherche à construire une fonction de prédiction globale f_h telle que $\hat{x}_{i,t+h} = f_h(x_{i,t}, \theta_{i,t})$. Cette approche est très populaire, car elle est une adaptation naturelle du cadre des régressions en Machine Learning. La difficulté est alors de choisir une bonne fonction f_h. On présentera un exemple de méthodes au chapitre 5.

— L'approche hiérarchique : L'idée est de se servir de la hiérarchie de produits dont on dispose, et d'effectuer ce qu'on appelle une réconciliation hiérarchique [SAH+21]. Le but est d'effectuer une prédiction à plusieurs niveaux de la hiérarchie, puis de rendre ces prédictions cohérentes entre elles.

— L'approche bayésienne : Cette approche consiste à essayer de prédire les ventes non comme une unique valeur, mais comme une distribution de valeurs plausible, ce qui a l'avantage d'aider à la création de stock de sécurité car on a alors une idée de la dispersion. Bien que cela ait été tenté au cours de la thèse, il n'a pas été possible de calibrer des prévisions intéressantes avec les données dont je disposais. D'une manière générale, les algorithmes d'optimisation bayésienne ne convergeaient pas, et les distributions obtenues étaient très étalées. Cependant, cette approche reste appréciée. Citons [Cha14] qui l'utilise pour prédire des séries de ventes pour du commerce physique.

4.5.4 Classification des features

. Il est donc impossible d'utiliser ces information pour prédire le futur. Cependant, il est possible de dépolluer les données passées en utilisant ces informations, afin de retraiter les données.

Au cours de la thèse , on a pu utiliser différent types de données externes. Ces données peuvent être classées de différentes façons.

Dimension de distribution des variables externes Tout d'abord, on peut classer les features selon la dimension (longitudinales ou temporelles selon laquelle elles sont distribuées)

— **Variables temporelles :** Covariables dépendant de la date t uniquement. Elles sont communes à tous les produits. Par exemple, les événements spéciaux (Noël, Black Friday) et les covariables liées aux conditions météorologiques entrent dans cette catégorie.

— **Variables longitudinales :** Covariables dépendant du produit i uniquement. Par exemple, le type de produit, sa marque. Les caractéristiques longitudinales permettent de creer une hiérarchie entre les produits.

— **Variables mixtes** : Covariables dépendantes des deux. Par exemple, le prix d'un produit peut varier chaque semaine.

Séparer les features ainsi permet de les appliquer à différents niveaux prédictions, par exemple pour des réconcialiation hierarchique. Ainsi, des variables temporelles peuvent être appliquées à des prévisions à la catégorie, alors que les variables longitudinales doivent être appliquées à la prédiction produit.

Prédictibilité des variables externes Une autre notion importante est la notion de prédictibilité des variables externes. Typiquement, on pose la question : "Est ce que je connais la valeur de cette features dans le futur à la date pour laquelle je veux faire la prévision ? ". On peut distinguer deux type de features :

— **Features imprévisibles :** Features imprévisibles pour l'horizon de prédiction. Un exemple immédiat est la météo, qui est difficilement connue au delà de 2 semaines. Plus ennuyeux, les prix sont généralement imprévisible. S'il est possible d'avoir une idée du niveau de prix futur d'un produit (CDiscount ne va pas vendre des Iphone à 1 €), connaître les prix exacts, et par extension la position des différents produits en terme de prix est impossible. Cela est du à des mécanismes de pricing qui peuvent dépendre de ce que vont faire les concurrents, de la stratégie commerciale future, de négociations avec les fournisseurs, etc....

— **Features prévisibles :** Features connues pour l'horizon de prédiction prévision, comme les évènement commerciaux, les caractéristiques produits comme la marque, le positionnement sur le marché, ... On peut y ajouter des caractéristiques produits qui ne sont pas connues exactement, mais qui évoluent peu, comme l'avis client moyen.

Il est impossible d'utiliser directement les features imprévisibles pour prédire le futur. Cependant, il est possible de dépolluer les données passées en utilisant ces informations, afin de retraiter les données.

4.6 Conclusion et Contributions des chapitres suivants

Faire une bonne prédiction des ventes est donc un problème complexe. Dans les chapitres suivants, on présentera des méthodes résolvant le seul problème (4.1), évaluées en utilisant des métriques classiques. Cependant, il faut garder en tête que ces métriques, et ce type de problème ne peuvent pas être le seul critère à l'aune duquel on évalue une prévision des ventes. Il est relativement facile d'optimiser une métrique donnée sans prendre en compte les autres critères, notamment l'interprétabilité et même la pertinence de cette métrique. C'est pourquoi dans le chapitre 5,on insistera également sur d'autres aspects que les seules métriques.

4.6.1 Contribution du chapitre 5

Le chapitre 5 propose une méthode de prévision auto-régressive de prévision des ventes. Au cours de la thèse, une variante de cette méthode, nommée *ARBoost* a été développée et mise en production pour la prévision des ventes à CDiscount. Le modèle proposé se base sur le cadre *auto-régressif* présenté au chapitre précédent, et utilise la puissance des algorithmes de boosting d'arbres pour la prévision des ventes. Si cette approche est relativement classique, l'originalité est ici la construction d'un certain nombre de features du modèle, comme un facteur de saisonalité comparable entre produits, y compris lorsque que ces produits présentent peu d'historique. On peut ainsi effectuer des prévisions "à froid", en disposant de très peu de données historisées sur les ventes des produits que l'on considère.

Le chapitre propose également un cadre de pré-traitement pour surmonter les difficultés inhérentes aux données de commerce électronique, comme par exemple les ruptures de stock, les outliers, et les features "imprévisibles".

Notons que la soumission dont est issu le chapitre 5 est relativement ancienne (2019), et que certaines conceptions présentés dans ce chapitre en sont absentes. Ainsi, la discussion sur les métriques de la section 4.3 n'ont pas été prise en compte, et on ne cherche dans cet article qu'à optimiser des métriques classiques (MAPE et MAE), sans se soucier de l'impact réel de ces métriques sur la prédiction des ventes. Par ailleurs, l'aspect interprétable est très largement absent de ce type de modèle "boite-noire".

ARBoost corrige nombre de ces défauts, en se concentrant sur des métriques plus adaptées, notamment le %MAED présenté plus tôt. Par ailleurs, on a cherché à développer des outils permettant de comprendre le niveau d'une prévision réalisé par ARBoost en servant notamment du package SHAP en Python[LL17]. Ce package se base sur l'utilisation de la valeur de Shapley issue de la théorie des jeux pour distribuer des gains à plusieurs joueurs ayant coopéré.

4.6.2 Contributions du chapitre 6

Ce chapitre propose une modélisation de la compétition entre différentes séries temporelles, avec une application à la prédiction des ventes en présence d'une cannibalisation forte. L'idée est de modéliser les ventes $x_{i,t}$ pour tous les produits d'une même catégorie par un processus analogue au processus auto-régressif pour séries catégorielles avec variables exogènes proposé par [FT17], mais utilisant un simple ratio au lieu d'une fonction logistique. On résume le modèle ainsi :

$$\begin{cases} x_{i,t} &= \epsilon_{i,t}(\lambda_{i,t}) \\ \lambda_{i,t} &= s(t)\dfrac{\phi(x_{i,t-1}/s(t-1),\theta_{i,t})}{1+\sum\limits_{j\in[1,d]}\phi(x_{j,t-1}/s(t-1),\theta_{j,t})} \end{cases}$$

Où :

— Les $\epsilon_{i,t}$ sont des processus de comptage (des processus de Poisson) qui modélisent la dispersion statistique autour d'une valeur $\lambda_{i,t}$,

— $s(t)$ est un paramètre qui estime la somme de toutes les ventes à une date t,

— ϕ est une fonction non-linéaire appliquée aux parts de marché précédentes $(x_{j,t-1}/s(t-1))$, et aux paramètres externes connus $\theta_{i,t}$.

Ainsi, il s'agit d'un modèle hiérarchique qui distribue la valeur agrégée $s(t)$ en fonction des différents paramètres $\lambda_{i,t}$ qui expriment la "compétitivité" d'un produit i à une date t. On "explique" les variations de part de marché en se servant des covariables (prix, marge). Cette approche permet de modéliser le fonctionnement de la cannibalisation, et semble améliorer les performances prédictives de modèle "classiques" lorsqu'on cherche ϕ sous la fonction d'un réseau de neurones simple.

Chapitre 5

Un cadre classique de prédiction appliqué au E-Commerce

Ce chapitre est issu d'une contribution à la conférence APIA 2019 (Applications Pratique de l'Intelligence Artificielle), parue initialement sous le titre : Une approche multi-séries pour la prévision de la demande sur des données d'E-Commerce. Une version similaire de cet algorithme, nommée ARBoost a été implémentée dans le cadre de la thèse et est actuellement utilisée à CDiscount.

5.1 Introduction

Le E-Commerce repose sur des prévisions opérationnelles de la demande au niveau des produits. En effet, les standard modernes des chaînes logistiques modernes exigent un réapprovisionnement *à flux tendus*, pour diminuer les coûts de stockage et les invendus. Une meilleure précision des prévisions peut entraîner d'importantes économies et une réduction du coût et de la place nécessaire au stockage.

Cependant, l'environnement commercial dans le commerce électronique rend cette prévision complexe en raison de la volatilité des ventes. Par exemple, les ventes sont affectées par le calendrier (jour fériés, vacances), les comportements des concurrents, les modifications de prix, *etc* .... Les données relatives à la demande comportent divers défis, tels que des données historiques non stationnaires, des séries chronologiques courtes et des effets de cannibalisation entre produits similaires.

On dispose généralement d'un arbre hiérarchique naturel entre produits, selon le type de produits. Cet arbre se constitue à l'aide d'informations sur le famille, le type de produit, la marque, etc... Dans cet arbre hiérarchique, des produits proches auront des comportements similaires. Ainsi, les produits de la famille 'Jouets' auront une saisonnalité similaire et connaîtront souvent leurs meilleures ventes de l'année avant Noël.

Les méthodes existantes traitent généralement différentes séries séparément. Cela fonctionne dans la vente au détail physique, mais la rotation rapide des produits et la volatilité de la demande dans la vente en ligne nécessitent de fournir des modèles qui partagent les informations entre les séries chronologiques[Yel10; Cha14; TKF15; BSB+19].

Dans cette étude, nous proposerons un cadre pour le problème de prévision de la demande du monde réel dans le commerce électronique. Notre objectif est d'exploiter la corrélation entre les séries pour améliorer la précision des prévisions. En particulier, nous cherchons à surmonter le problème de la faible longueur des séries chronologiques.

Dans la section 5.2, nous définissons formellement le problème et proposons un bilan rapide des travaux antérieurs sur le terrain. Nous présentons un pré-traitement des données dans la section 5.3. Dans la section 5.4, nous présentons le modèle de boosting qui nous donne les meilleures performances. Enfin, nous présentons la configuration et les résultats de notre expérience sur un ensemble de données réelles sur la section 5.5.

5.2 Contexte

5.2.1 Position du problème

Nous avons un ensemble I de produits, divisé en K différentes catégories I_k telles que $I = \biguplus_k I_k$ (I union disjointe de I_k). Nous avons également N séries $(y_{i,t})$, où $y_{i,t}$ représente le nombre de ventes du produit $i \in I$ au cours de la semaine t. Cette série s'observe pendant les T semaines.

Le support de cette série, c'est-à-dire le nombre de semaines non nulles pour chaque série est relativement faible par rapport à T. Cela signifie que nous n'observons pas beaucoup d'historique pour chaque produit individuellement. Nous supposons que les séries suivent une saisonnalité de période τ, la plupart du temps annuelle ($\tau = 52$), bien qu'une période entière soit rarement observée pour un produit donnée.

On notera $Z = ((z_{i,t})_i)_t$ les features présentées en section 4.5.4.

Notre objectif est de prévoir les valeurs de cette série pour un horizon h. Plus formellement, nous souhaitons développer un modèle de prévision f, tel que, si nous considérons les ventes passées d'un produit i $y_{i,:t} = (y_{i,0}, \ldots, y_{i,t})$, la valeur $f(y_{i,:t}, z_{i,t}, \theta)$ est un estimateur de $y_{i,t+h}$ pour un ensemble de paramètres θ pouvant être appris .

5.2.2 Travaux Similaires

Un grand nombre de travaux ont été publiés concernant les méthodes de prévision de la demande, pour différentes applications (installations, vente au détail physique et en ligne, ...). Les méthodes les plus largement utilisées sont les modèles de séries chronologiques classiques tels que les modèles ARIMA [EA07] et les variantes de lissage exponentiel [Tay03]. Cependant, les prévisions dans le domaine du commerce électronique doivent généralement faire face à des problèmes tels que les tendances des ventes irrégulières, la présence de données de vente extrêmement volumineuses et éparses, *etc.* Certaines de ces limitations peuvent être surmontées grâce à la fonction de vraisemblance modifiée et aux modèles linéaires étendus [SSF16]. Mais cette méthode ne parvient pas à obtenir de bonnes performances lorsque les séries sont petites.

D'autres méthodes de régression ont été proposées. Par exemple [PG11] utilisent des modèles additif généralisées pour la demande d'électricité, [CC+04] utilise,t des SVR(régression à support de vecteur) et [BBO17] des réseaux de neurones récurrents. Toute ces méthodes ne croisent pas les informations disponibles entre séries et s'étendent mal au problème de prévision de la demande pour le E-commerce.

Récemment, [BSB+19] ont proposé d'utiliser des réseaux de neurones profonds pour réaliser des prédictions en transférant des informations entre séries dans le cadre du E-commerce. Ils adaptent une architecture de réseau de neurones(Long Short Term Memory ou LSTM) pour traiter toutes les séries en même temps. Ils séparent également les effets des caractéristiques longitudinales et temporelles pour obtenir de bonnes performances. Cela suggère que le partage d'informations entre séries permet d'améliorer les performances en prédiction.

Les modèles hiérarchiques bayésiens sont un autre modèle prometteur [Yel10; Cha14]. Ces modèles expriment les ventes d'un produit comme issues d'une distribution dont les paramètres sont eux même issue d'une distribution dépendant des caractéristiques du produits (prix, type). On a donc des équations à plusieurs niveaux, à la fois au niveau du produit et au niveau de sa catégorie. Ces modèles permettent d'établir des relations plus explicites entre les prédictions et les features, ainsi que de donner des bornes pour les intervalles de confiance.

5.3 Pré-traitement des données

5.3.1 Données de Ventes

Il existe deux types de problèmes avec les données de vente dans le commerce électronique. Le premier est la présence de valeurs anormalement basses, ou "faux zéros". Ces faibles valeurs

peuvent être dues à des ruptures de stock, à des problèmes de réseau ou à la modification du moteur de recherche sur le site Web. Notre objectif est de prédire une demande, qui a pu se reporter sur autre chose durant les semaines anormales (autre produits, concurrents,...) .Nous devons donc identifier et remplacer ces valeurs par des valeurs 'raisonnables' pour la demande. La correction des faux zéros se fait en remplaçant les valeurs anormalement 'basses' par des valeurs fictives issues d'algorithme univariés simples sur chaque séries temporelles . Ces valeurs serviront à entraîner les modèles, mais ne seront pas utilisées pour l'évaluation.

Le deuxième problème est la présence de valeurs anormalement élevées. Ces valeurs sont informatives, car elles nous renseignent sur l'effet des ventes. Cependant, ces valeurs sont problématiques lorsqu'elles sont utilisées en tant que variables explicatives, car elles peuvent suggérer un niveau de ventes supérieur aux prévisions ou donner des informations erronées sur les tendances et la saisonnalité. Par conséquent, nous construisons des 'ventes lissées' $x_{i,t}$ en éliminant les valeurs supérieures à γ multipliées par la variation standard. Plus précisément :

— Pour chaque produit, nous calculons une moyenne et un écart-type mobile

$$\overline{y_{i,t}} = \frac{1}{M}\sum_{k=0}^{M} y_{i,t-k}$$

$$\overline{\sigma_{i,t}} = \left(\frac{1}{M}\sum_{k=0}^{M}(y_{i,t-k} - \overline{y_{i,t}})^2\right)^{\frac{1}{2}}$$

.

— Si $y_{i,t} > \overline{y_{i,t}} + \gamma\overline{\sigma_{i,t}}$, alors $x_{i,t} = \overline{y_{i,t}} + \gamma\overline{\sigma_{i,t}}$.

— Sinon $x_{i,t} = y_{i,t}$.

5.3.2 Features temporelles : Saisonnalité et Tendance

La saisonnalité et la tendance d'une série temporelle sont deux caractéristiques essentiellement temporelles, et qui peuvent difficilement être déduites par un algorithme de machine learning utilisé en régression, comme ce qui sera présenté en section 5.4. Afin d'enrichir notre apprentissage, et d'introduire des aspects temporels dans notre régression, nous allons construire des features temporelles dans notre modèle correspondant à des tendances locales, des tendances annuelles et à des saisonalités pour chaque produit.

Construire une tendance locale est relativement évident. Etant donné un produit i, on peut calculer une tendance locale par régression (linéaire) sur une fenêtre glissante le long de la série. Cependant, il vaut mieux prendre une fênêtre assez large pour ne pas tenir compte des mouvemens les plus violents. Il est possible d'effectuer des calculs de saisonnalités annuelles de manière analogue lorsque l'information est disponible. Lorsque ce n'est pas le cas, on peut calculer une saisonnalité annuelle unique pour chaque catégorie de produits, représentant l'évolution annuelle du marché pour un produit particulier.

Le traitement des saisonnalités est plus complexe en raison de la brièveté des séries considérées. Nous utilisons une variante de la procédure décrite dans [KPW02] pour produire un facteur de saisonnalité pour chaque produit. Esquissons cette procédure.

Tout d'abord, nous normalisons les chiffres de vente pour chaque année. Nous voulons nous assurer que chaque produit a le même niveau moyen. Pour chaque produit i, en considérant N_i le nombre de semaines de vente au cours de l'année, nous notons pour une date t cette année (c.-à-d. $T \in 0,\dots,\tau-1$) :

$$x_{t,i}^{std} = \frac{N_i}{\tau}\cdot\frac{x_{t,i}}{\sum_{t=0}^{\tau} x_{t,i}}$$

Deuxièmement, nous calculons la moyenne des valeurs normalisées dans chaque catégorie I_k. Nous avons donc une saisonnalité standardisée pour chaque catégorie de produit. L'idée centrale

est de supposer qu'il existe une saisonnalité multiplicative commune $s_{I_k}(t)$ pour tous les produits de cette catégorie. Par conséquent, si la date à laquelle le produit a été mis sur le marché est uniformément répartie, la moyenne calculée est directement proportionnelle à la saisonnalité.

Toutefois, en raison de la nature erratique des données sur les ventes dans le E-commerce, à ce stade, la saisonnalité calculée n'est souvent pas assez informative et est souvent très bruité par la présence d'évènement particuliers propres à une année et une catégorie de produits (par exemple, les coupes du monde de football pour les ventes de téléviseurs) .

C'est pourquoi nous utilisons un algorithme de clustering de séries chronologiques pour regrouper les saisonnalité des différentes catégories. Ce regroupement est basé sur la distance euclidienne entre les modèles de saisonnalité, mais prend également en compte la variabilité des saisonnalités constaté dans chaque catégories.

Après clustering, nous obtenons un faible nombre de pattern de saisonnalités normalisées différentes, que nous introduisons comme variables explicatives features dans l'algorithme de machine learning.

5.3.3 Traitement des autres features

Traitement des features catégorielles Le E-Commerce dispose naturellement de nombreuses variables longitudinales catégorielles sur les produits (famille de produits, type, marque,gamme de prix, avis clients,...) qu'on aimerait pouvoir exploiter. Plusieurs méthodes sont possibles :

Tout d'abord, on peut séparer les produits par catégories, et entraîner un modèle pour chaque catégorie de produits. Cependant, on se heurte rapidement à la multiplicité des catégories considérées. Cela impose d'utiliser des algorithmes très simples, et empêche de capturer des effets plus faibles qui apparaîtraient en considérant un plus grand nombre de produits. Dans les exemples, on distinguera un modèle 'Par Magasin', qui entraîne un algorithme de Machine Learning par magasin, et un modèle global, qui entraîne un modèle sur l'ensemble du site.

Pour traiter des features plus complexes, on est donc amener à encoder les features catégorielles, c'est à dire à coder numériquement chaque catégorie. La méthode classique, dite de One-Hot Encoding, consiste à introduire une colonne booléenne pour catégorie. Cependant, du fait du grand nombre de catégories, cette méthode n'est pas praticable en pratique.

Deux possibilités subsistent. Tout d'abord, il est possible d'encoder les variables catégorielles sur plusieurs colonnes via une fonction de hashage. Cette méthode limite l'espace nécessaire pour l'apprentissage, mais rend plus difficile l'interprétation de l'importance relative des différentes variables.

Une autre possibilité consiste à utiliser un encoding ordinal, c'est à dire à associer un entier à chaque catégorie. Cette méthode est extrêmement simple, et permet une interprétation facile des résultats, mais introduit un ordre sur les features qui ne repose sur rien. Pour limiter cet effet, on peut agréger les résultats issus de plusieurs permutations possibles des features catégorielles.

Traitement des features imprévisibles Certaines featuress mixtes ou temporelles, telles que la météo ou les prix, ne peuvent pas être utilisées pour la prédiction, car elles ne peuvent pas être prédites pour l'horizon sur lequel nous voulons prévoir des informations. Cependant, ces fonctions peuvent être utilisées pour former le modèle sur les données passées, afin d'expliquer des valeurs anormalement basses (ou élevées) dans le passé. Nous pouvons ensuite effectuer une prédiction en utilisant une estimation des valeurs futures. Par exemple, nous pouvons prendre la valeur saisonnière des données météorologiques ou le prix passé moyen observé des produits considérés . Ce schéma a une faiblesse : le fait que nous utilisions des valeurs antérieures exactes conduit l'algorithme d'apprentissage automatique à donner beaucoup d'importance à ces features.

5.4 Modèle

5.4.1 Schéma d'apprentissage

Nous considérons notre problème de prévision de séries chronologiques multiples comme un problème de régression. Notre objectif est les valeurs de vente corrigées des "faux zéros" à l'horizon $y_{i,t+h}$. Nous utilisons les valeurs passées des ventes lissées comme des features, comme décrit dans 5.3.1. Nous avons donc une prédiction

$$\widehat{y_{i,t+h}} = f(x_{i,:t}, z_{t+h,i}, \theta)$$

L'hypothèse est que $x_{i,t}$ représente le niveau de vente 'normal'. Il est supposé supprimer les effets des effets ponctuels, comme les offres spéciales. Les variables externes $z_{i,t}$ nous donnent des informations sur la différence $\delta_{i,t} = y_{i,t} - x_{i,t}$. Par conséquent, nous préférons utiliser les valeurs décalées de la variable lissée $x_{i,t}$ comme variable explicatives plutôt que $y_{i,t}$.

Nous avons utilisé comme ensemble d'apprentissage les valeurs des tuples $(x_{i,:t}, z_{t+h,i})$ pour tous les produits i avant une date donnée. Les hyper-paramètres sont sélectionnés en utilisant une simple période de validation.

Nous résumons tout sur la figure 5.1.

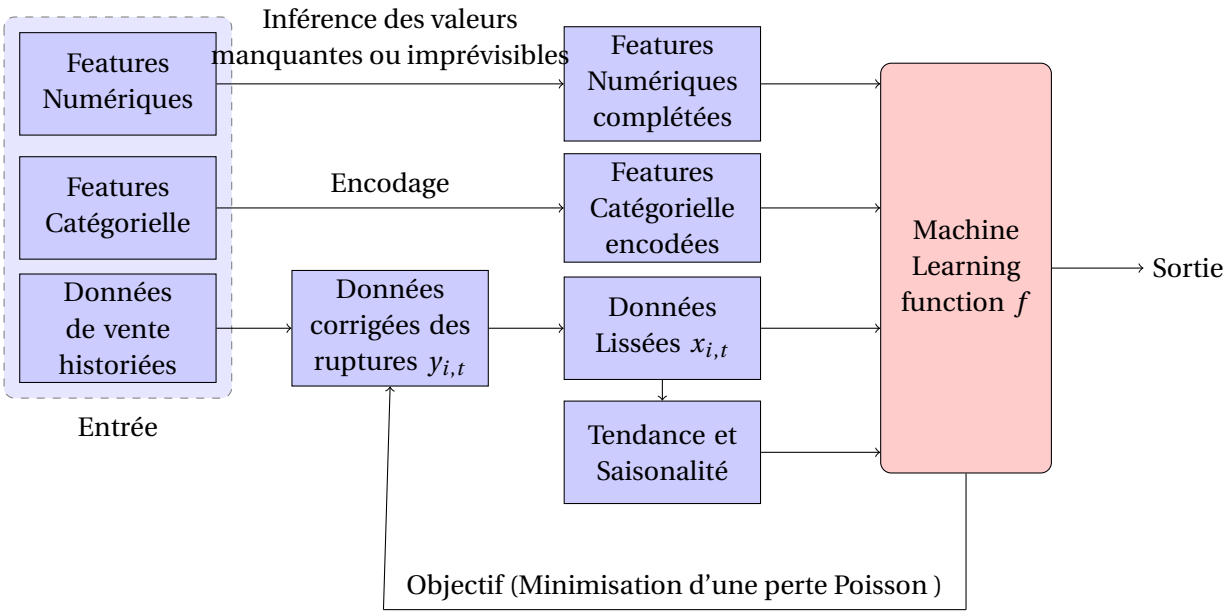

FIGURE 5.1 – Schéma général du fonctionnement de l'algorithme de prédiction

5.4.2 Métriques d'évaluation et métriques d'apprentissages

Nous devons différencier les métriques d'évaluation des métriques utilisées lors de l'apprentissage de la fonction de Machine Learning f. Nous avons utilisé comme métriques d'évaluation de l'erreur quadratique moyenne (RMSE) et de l'erreur absolue moyenne (MAE) en tant que métriques d'évaluation, en utilisant le prix du produit p_i comme poids. [1] Les prix considérées sont des Prix unitaire moyen pondéré, qui représente la valeur à l'achat d'un produit du stock. Pour les besoins de l'évaluation, cette valeur est constante.

$$\text{RMSE}(f, \theta) = \sqrt{\frac{1}{n} \sum_{i \in I} p_i^2 \cdot (y_{i,t+h} - f(x_{i,:t}, z_{t+h,i}, \theta))^2}$$

1. Cette introduction du prix est issue de l'article original, et n'a pas pu être corrigé pour être consistante avec le reste de la thèse, ou l'on parle en nombre de produits. On exprimera donc les RMSE en euro dans ce chapitre.

$$\text{MAE}(f,\theta) = \frac{\sum_{i\in I} p_i \cdot |y_{i,t+h} - f(x_{i,:t}, z_{t+h,i}, \theta)|}{\sum_{i\in I} p_i \cdot f(x_{i,:t}, z_{t+h,i}, \theta)}$$

Ces métriques sont couramment utilisées dans les prévisions de chaîne d'approvisionnement. Cependant, ces métriques souffrent de plusieurs faiblesse pour les utiliser comme métriques d'apprentissage. En effet, le RMSE et le MAE ont tendance à sous-estimer la prévision en raison du caractère positif de la série. De plus, l'erreur de prédiction des produits les plus vendus a tendance à l'emporter sur les autres erreurs.

On va utiliser une métrique différente pour modéliser la dispersion de la série. On va en effet supposer que, les valeurs $y_{i,t}$ sont tirées par des distributions de Poisson indépendantes du paramètre $lambda_{i,t}$ pour tout produit i et toute date t.

Ce choix a déjà été utilisé, par exemple par [BBO17] et est naturel pour trois raisons. Premièrement, nous avons observé que la variance des séries est proportionnelle à la moyenne empirique des séries, avec des coefficients de proportionnalité proche de 1, ce qui est ce qu'on attend dans le cas d'une distribution de Poisson.
Deuxièmement, cela nous permet de limiter les effets de la présence de valeurs aberrantes dans nos données. En effet, des valeurs plus élevées sont plus probable que dans une modélisation par un bruit blanc gaussien, par exemple.
Troisièmement, les valeurs entières positives sont naturellement modélisées par un processus de comptage. Nous pouvons supposer que pour chaque semaine t et chaque produit i, les dates d'arrivées du client suivent un processus de Poisson et que le paramètre de ce processus change chaque semaine.

C'est pourquoi on peut utiliser une perte de Poisson lors de la phase d'apprentissage de notre modèle . Si l'on suppose que $y_{i,t}$ est distribué selon une distribution de Poisson de paramètre $\lambda_{i,t}(\theta) = f(x_{i,:t}, z_{t+h,i}, \theta)$, le critère que nous voulons optimiser est alors la log-vraisemblance de la valeur $y_{i,t}$ en faisant varier les paramètres θ de la fonction d'apprentissage :

$$\text{Poisson}(f,\theta) = \sum_{i,t} \lambda_{i,t}(\theta) - y_{i,t}\log(\lambda_{i,t}(\theta))$$

5.4.3 Algorithme

Le choix de l'algorithme d'apprentissage automatique pour calculer f et θ est crucial. D'une part, il doit être suffisamment souple pour utiliser différents types de features et pour sélectionner les variables explicatives les plus utiles. En particulier, il devrait pouvoir résister à une redondance des features . En revanche, il doit être suffisamment consistant pour éviter le sur-apprentissage. Enfin, en raison du grand nombre de séries et de features, il doit être suffisamment rapide pour gérer des données volumineuses.

Nous avons testé différents modèles. Pour chacun, nous avons sélectionné les hyper-paramètres par validation croisée. Nous essayons également de normaliser les features dans les différents cas.

Modèles linéaires Le choix le plus simple est de chercher une fonction f linéaire. Dans ce cas, la simplicité du modèle autorise même d'utiliser le One-Hot Encoding. Il est également possible d'ajouter une régularisation L1 ou L2 pour éviter l'overfitting. Cependant, les effets étudiés semblent profondément non-linéaires, et dépendent d'effets croisés entre features que les modèle linéaires ne prennent pas en compte.

Modèle additifs généralisées (GAM) On peut étendre un peu les modèle linéaires, et utiliser des modèles linéaires généralisées(GAM). L'idée est de chercher la fonction f dans une base de fonctions relativement simples. Les modèles GAM sont fréquemment utilisés pour prédire des séries temporelles(voir [Has17]) et peuvent prendre en compte des effets croisés. Cependant, il n'est pas évident de choisir une bonne base de fonction f, et nous ne sommes pas parvenu à trouver une bonne modélisation par ces modèles.

Forêts aléatoires Les forêts aléatoires sont un type d'algorithme de bagging, qui consiste à construire différents arbres de régression par boostrap, puis à produire une prédiction basée sur les prédictions des différents arbres. Il permet de prendre en compte les effets de seuil et des effets croisés. Il peut être parallélisé, ce qui permet un calcul rapide.

Les forêts aléatoires conviennent bien à l'estimation de f et permettent donc d'obtenir de bonnes performances sur les jeux de données.

Arbres boostés Contrairement aux méthodes de forets aléaires, les méthodes d'arbres boostées implémentent un regroupement séquentiel de la prédiction de différents arbres. Ils ont récemment fait l'objet de beaucoup d'attention, en raison de leurs performances sur des cas réels. Ici, nous utilisons principalement XgBoost [CG16], qui en est une implémentation à gradient rapide.

Il conserve les avantages des forêts aléatoires, mais offre de meilleures performances. Le prix est généralement un temps de formation plus long, car la formation ne peut pas être mise en parallèle. Les hyper-paramètres XgBoost sont sélectionnés via validation. Les domaines de validation des hyper-paramètres sont présentés sur la table 5.1. Nous utilisons un arrêt précoce pour réduire le temps d'entraînement.

hyper-paramètres	Min value	Max value
learning rate	0.01	0.3
min split loss	0.01	0.2
max depth	5	8
round evaluation	1000	5000

TABLEAU 5.1 – Domaine utilisées pour rechercher les hyper-paramètres pour XgBoost

5.5 Expériences

5.5.1 Description du jeux de données

Pout tester le cadre que nous proposons pour la prévision des ventes, nous allons prendre un ensemble de données provenant de *Cdiscount.com*. Il rassemble les ventes de 99305 produits, répartis dans 10 magasins et 1031 catégories, sur une période d'environ 4 ans. Sur la figure 5.2, nous avons représenté la répartition des produits par durée de vie. Seule une minorité des produits depasse les 52 semaines de ventes.

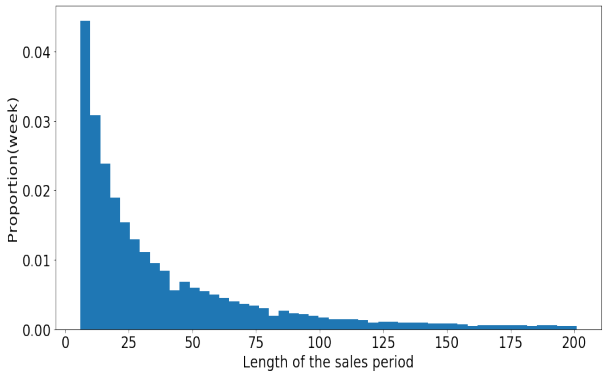

FIGURE 5.2 – Répartition des produits par durée de vie

En figure 5.3 quelques exemples de courbes de ventes pour des téléviseurs.

Nous prenons un horizon h de 6 semaines, puis formons le modèle sur les 170 premières semaines, puis utilisons les 10 prochaines semaines pour valider les hyper-paramètres. Pour évaluer

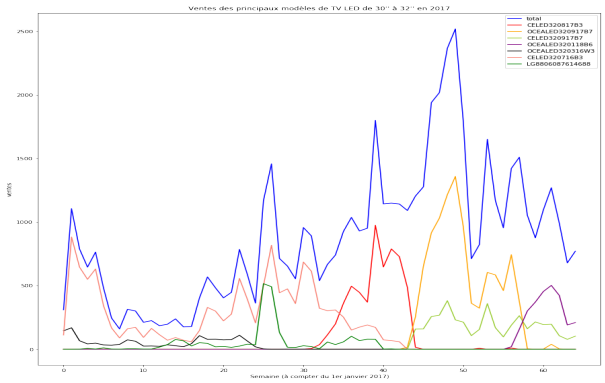

FIGURE 5.3 – Ventes de quelques modèles de téléviseurs

le modèle, nous utilisons les 19 dernières semaines. Ces semaines correspondent aux dernières semaines de l'année 2018 et au début de 2019. Elles contiennent donc beaucoup de variabilité (Black Friday, Noël, soldes d'hiver).

Il existe 3 ensembles de produits, nommés A, B et C. Le premier regroupent les produits qui se vendent le plus, le dernier les produits qui se vendent le moins.

5.5.2 Benchmark et variantes

Nous comparons notre approche à des algorithmes "maison", ainsi qu'à des solution industrielles.

Plus précisément, nous utilisons un simple algorithme de lissage exponentiel comme référence (ES). Par ailleurs nous comparons avec un algorithme commercialisé, développé par la société Relex[Rel]. Cet algorithme effectue une prédiction pour chaque série en utilisant une classification des séries chronologiques et des connaissances métier.

Nous présentons les performances de l'algorithme XgBoost dans différentes configurations. Nous faisons varier l'encodage des variables catégorielles, ainsi que l'ensemble de données considérées. Dans le cadre global, nous entraînons l'algorithme sur l'ensemble des produits disponibles. Dans le cadre mixte, on entraîne un faible nombre (4) de magasins très particuliers séparément. Enfin, on étudie l'impact de l'ajout d'une feature modélisant la saisonnalité sur nos prédictions.

Un avantage de la méthode de prédiction globale est qu'elle permet la prédiction *à froid*, c'est-à-dire la prédiction de nouvelles séries sans historique. Pour obtenir une évaluation similaire à celle du benchmark, nous ignorons les 6 premières semaines de vie des produits, où notre algorithme est capable de prédire, mais pas les algorithmes de référence.

5.5.3 Résultats

Le tableau 5.4 présente les performances de la prévision pour deux mesures d'évaluation pour l'ensemble des produits et pour les différents ensembles A, B et C. Les valeurs RMSE sont exprimées en milliers d'euros (k€). Nous présentons différentes versions de notre algorithme en fonction du codage des caractéristiques catégorielles (ordinal ou hachage) et de l'utilisation d'une feature exprimant la saisonnalité (décrites en section 5.3.2).

Nous pouvons voir que XgBoost surpasse le benchmark pour toutes les catégories. Il réduit le MAE d'environ 5% et le RMSE de 10% sur l'ensemble des jeux de données. Globalement, le gain relatif est plus important dans le RMSE que dans le MAE, ce qui montre qu'il réduit généralement le plus grand écart de performances plus qu'il n'améliore la prédiction moyenne.

L'introduction d'une variable modélisant la saisonnalité améliore les performances en général pour les produits les plus vendus, notamment en RMSE. Cependant, elles semblent dégrader les performances sur les produits les moins vendus. On peut supposer que l'ajout d'une saisonnalité n'est vraiment sensible qu'à partir d'un certain niveau de ventes.

Bizarrement, l'encodage ordinal semble plus efficace que le hashing.

	Cadre			**Tous**		**A**		**B**		**C**	
ML Algo.	Configuration	Encodage	Méthode	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
ES				3.83	1.09	5.68	1.03	1.12	1.06	1.28	1.31
RF	with seas.	Ordinal	Global	3.09	0.831	5.27	0.796	1.66	0.892	1.32	0.92
XgBoost	Poisson/avec seas.	Ordinal	Global	**2.67**	0.725	**4.59**	0.674	1.41	0.801	**1.20**	0.874
XgBoost	Poisson/sans seas.	Ordinal	Global	2.76	0.730	4.72	0.681	1.39	0.800	1.21	0.872
XgBoost	Poisson/avec seas.	Hashing	Global	2.78	0.728	4.75	0.685	1.41	0.816	1.21	0.893
XgBoost	Poisson/sans seas.	Hashing	Global	2.79	0.740	4.77	0.689	1.40	0.817	1.22	0.893
XgBoost	Poisson/avec seas.	Ordinal	Mixte	2.80	**0.707**	4.67	**0.656**	**1.33**	0.788	1.23	0.852
XgBoost	Poisson/sans seas.	Ordinal	Mixte	2.79	**0.707**	4.77	0.658	1.36	**0.785**	1.25	**0.85**
Benchmark				3.01	0.758	4.97	0.688	1.77	0.907	1.56	0.982

TABLEAU 5.2 – Comparaison de différents modèles

Si nous examinons attentivement les performances, nous constatons que notre algorithme est particulièrement performant au cours des semaine du début de cycle du produit. Nous présentons ces résultats dans 5.3 pour le benchmark et un XgBoost avec saisonnalité . La forte variabilité de la RMSE est due au faible nombre de produits concernés et à la forte variabilité de la période étudiée. Néanmoins, nous pouvons constater que, au début du cycle de produit, notre structure dépasse fortement la référence. Nous diminuons par exemple le MAPE de 24,0 % et le RMSE de 42,5 % pour le produit avec 10 semaines de données historiques. Cette différence diminue avec le temps, à mesure que le benchmark acquiert un historique suffisant pour sa prédiction.

Product cycle	**Framework**		**Benchmark**	
Longueur	RMSE	MAE	RMSE	MAE
8	3.71	1.04	6.04	1.77
9	3.28	0.879	6.43	1.36
10	3.67	0.920	6.39	1.21
11	11.48	1.15	11.97	1.31
12	5.33	0.867	7.19	0.928

TABLEAU 5.3 – Performance sur le début du cycle de vie des produits

5.6 Conclusion

L'amélioration de la prévision de la demande dans le commerce électronique est possible grâce à l'utilisation de méthodes globales, qui partagent des informations entre des séries temporelles. Dans notre article, nous avons proposé d'utiliser une méthode de renforcement de gradient pour le faire, mais d'autres méthodes sont en développement[BSB+19]. Cela nous permet d'exploiter les caractéristiques croisées et les effets non linéaires existant dans les données de commerce électronique. En outre, nous pouvons également effectuer la prédiction *à froid*, avec très peu d'historique sur nos produits.

Nous avons également proposé plusieurs astuces pour résoudre les difficultés inhérentes aux données de commerce électronique. En particulier, nous avons proposé un moyen de calculer la saisonnalité des produits grâce au comportement des catégories de produits considérées.

Enfin, nous évaluons notre méthodologie sur un ensemble de données du monde réel, avec un nombre de produits réaliste, et surpassons les solutions de pointe en matière de prévision de la demande.

Cependant, nous n'avons pas de modélisation de la compétition 'entre produits' au sein d'un site de E-commerce. D'autres travaux seront nécessaires pour développer des modèles pouvant prendre en compte cette compétition.

Chapitre 6

Modèle de compétition pour la prévision des ventes

Competition between times series often arises in sales prediction, when similar products are on sale on a marketplace. This article provides a model of the presence of cannibalization between times series. This model creates a "competitiveness" function that depends on external features such as price and margin. It also provides a theoretical guaranty on the error of the model under some reasonable conditions, and implement this model using a neural network to compute this competitiveness function. This implementation outperforms other traditional time series methods and classical neural networks for market share prediction on a real-world data set. Moreover, it allows controlling underprediction, which plagues traditional forecasts models.

6.1 Introduction

Forecasting multiple time series is a useful task, which has many applications in finance [TC01], load forecast [ZWC18], health care[Lin89], retail operations[RCS20] or in supply chain management [AQS01; Van21]. This is however a complex task because the number of possible interactions between times series grows likes the square of the number of time series. For simple models such as Vector Autoregression (VAR) [Lüt05], it means that the number of times series should remain small. That is why practitioners tend to use tricks like sparsity [DZZ16]. However, most of the time, they forecast each series individually (see classical models in [Van21] for instance), or build an auto-regressive framework treating each point of data in a similar way [BSB+19; GB19]. However, in both cases, these methods neglect the possible interaction between time series.

In this article, we are interested in a specific type of interaction between time series : cannibalization. Product cannibalization has been defined as “the process by which a new product gains sales by diverting sales from an existing product” [SRG05]. In this paper, we want to model and forecasts the sales of different products in presence of cannibalization. We observe multiple time series that represent similar assets (or products) that compete with each other. Our objective is to forecast the future values of this time series by taking cannibalization into account to improve the demand forecast. Therefore, we need to introduce external covariates (for instance, the prices of different products) that may explain the cannibalization.

We will apply this model to E-commerce sales data. These data are organized in a hierarchy of categories. For instance, in the family ’HOME’, there is a subfamily ’Home Appliance’ which contains a category ’Fridge’, which can also be further subdivided. It is generally easier to predict aggregated sales for a category than to predict the sales of each product in this category. One of the reasons is the competition between the different products, and the other cross products effects. For instance, the cheapest products and the best-ranked products in the research engine achieve a competitive advantage. However, these advantages do not last forever, with the introduction of new products on the markets. Furthermore, prices and ranking in search engines change every day. Therefore, the competitiveness of each product changes every time step.

In section 6.2, we present the model used to predict E-commerce sales. In section 6.3, we establish an oracle bound on the estimation risk of our model. In section 6.5, we present the application of our model on the various dataset provided by the French E-Commerce retailer *CDiscount.com*.

Previous Work

Managers are generally aware of the presence of cannibalization and competition between assets [BMW01], but there are few attempts to model and estimate the impact of cannibalization.

In particular, the very well-known [BLP95] proposes a model of the cross-price elasticity of different products in the U.S. automobile market. They consider the sales of the different models and use household data, such as the composition of the household, its income, its location to model the choice of the consumer. This information is aggregated at the geographic level. This work has been extended by [BLP04], which considers individual information on each client, instead of aggregated data. This type of models is often used in the automobile industry to forecast sales. It concerns however a large amount of data to be used and is generally used only for long-term prevision.

There has also been some work to identify the presence of cannibalization in a different context, for instance in the beverage industry [SRG05], or in presence of innovative products [VHSD10]. The interested reader may also refer to this last paper to have a more detailed overview of cannibalization identification.

The originality of the method proposed in this paper is to use a machine learning approach for modeling competition and to use external covariates to explain cannibalization. We do not use any information on the behavior of the consumer.

Notations. We set $\mathbb{N} = \{0, 1, 2, \ldots\}$ and we also also $\|x\| = \sum_{i=1}^{d} |x_i|$, if $x = (x_i)_{i\in[1,d]} \in \mathbb{R}^d$.

6.2 Model

6.2.1 Observations

We observe a multi-dimensional time series $X_t = (x_{i,t})_{i\in[1,d]}$ in $\mathbb{N}^d$. For sales prediction, the integer $x_{i,t}$ represents the sales of the product i at the date t. Within this set, products are supposed similar and in competition with one another. For instance, it could be a single type of product or different types of products having the same usage.

We have n observations of this time series. In many cases, n has the same order of magnitude or is smaller than d.

We suppose that we know a non-negative estimator $s(t)$ of the expected total amount of sales $\mathbb{E}[\sum_{i=1}^{d} x_{i,t}]$ at time t given those n observations. This estimator is independant of the actual $x_{i,t}$, and is supposed given before we observe the actual sales. Computing such an estimator is generally an easier task than the amount of sales for a specific product at time t, because it is always easier to predict aggregated values than to predict multiple values, and because the global behavior of the series is generally easier to predict than an individual one. Many classical uni-dimensional techniques could be used to compute such an estimator, and we will not discuss this aspect.

Method for prediction of aggregated sales can be found in [BJRL15; Van21]. It is also possible to cite [KP10] for the specific case of aggregated sales prediction.

Let $y_{i,t} = \frac{x_{i,t}}{\sum_i x_{i,t}}$. It is the *market share* of the product i at the date t. However, this market sahre are not easy to manipulate from a theoretical perspective. Therefore, we will instead consider in this section an approached value of the market share $\widetilde{y_{i,t}} = \frac{x_{i,t}}{\sum_i s(t)}$.

We also observe a series of covariates $(\theta_{i,t}) \in \mathbb{R}^p$. These covariates are correlated to the values of the series $x_{i,t}$. In the sales forecasting setting, it could for instance represent the price or the profit margin of the product.

The forthcoming section will lead to a rigorous construction of the model.

6.2.2 Modeling dispersion

The first step is to model the dispersion of the series. It is natural to suppose that $x_{i,t}$ are drawn from a Poisson distribution with parameter $\lambda_{i,t}$.

More precisely, we suppose that there exists $(\epsilon_{i,t})$ independent unit Poisson processes such that $x_{i,t} = \epsilon_{i,t}(\lambda_{i,t})$. We need to introduce such process $(\epsilon_{i,t})$ in order to distinguish the "natural" stochastic dispersion of the random variable and the variation of the parameter $\lambda_{i,t}$. This distinction would be useful later, because, we hope to explain the variation of the parameter $\lambda_{i,t}$.

In the case of E-commerce sales forecasting, the choice of Poisson distributions has already been suggested in [BSB+19] : it has several advantages.

Firstly, we observed that the sales time series are strongly heteroscedastic and that the local variance of the series is strongly correlated with the local mean of the time series. This phenomenon also appears when the dispersion is modeleded by Poisson processes.

Secondly, it allows us to restrict the effects of the presence of outliers in our data. Indeed, higher values are more likely than with a Gaussian white noise modeling for instance.

Thirdly, positive integer values are naturally modeled by a counting process. We can suppose, that for each week t and each product i, the arrival of clients follows a Poisson process and that the parameter of this Process change for each time period t. It implies, that arrival time of the different clients are independent conditionally to the parameter $\lambda_{i,t}$.

We could also have use binomial negative distribution, in order to model over-dispersion. However, this would add another parameter to estimate and likely increase the risk of overfitting.

6.2.3 Modeling competition

Now, we want to model the competition and the cannibalization between the different time series. The main idea is to introduce weights for each product and each date. Such a weight represents the competitiveness of each product, which may vary over time. Then, we distribute the sales proportionally to this weight.

More formally, for each series i at each date t, we introduce a weight $w_{i,t}$. Then the parameters $\lambda_{i,t}$ in the previous section are defined as :

$$\lambda_{i,t} = s(t) \cdot \frac{w_{i,t}}{1 + \sum_{j=1}^{d} w_{j,t}}$$

The "+1" is here to ensure that the magnitude of the weight remains the same for all the observed periods. Therefore :

$$\mathbb{E}\left[\sum_{i=0}^{d} x_{i,t} | w_{1,t}, \dots w_{d,t}\right] = s(t) \cdot \frac{\sum_{i=0}^{d} w_{i,t}}{1 + \sum_{i=1}^{d} w_{i,t}}$$

If the sum of the weights is large enough, then $s(t)$ is a good estimator of the expected sum $\mathbb{E}\left[\sum_{i=0}^{d} x_{i,t} | w_{1,t}, \dots w_{d,t}\right]$, and in particular, if the sum of the weights are almost always large enough, this value is constantly close to $s(t)$ for all possible weights. In this case, $s(t)$ becomes also a good estimator of $\mathbb{E}\left[\sum_{i=0}^{d} x_{i,t}\right]$.

It is easy to add a new product in this setting just by adding a new weight. It is useful, because of the short sales cycle of numerous products.

6.2.4 Modeling temporal evolution

In this section, we explain how the weight of the previous section are computed and how they vary time. We suppose that there is a function ϕ, such that

$$w_{i,t} = \phi(\widetilde{y_{i,t}}, \theta_{i,t}) \tag{6.1}$$

. Let us explain the assumption behind this relation. To begin, we should note that the function ϕ is applied to two different parameters. The first one relies on the past values of the market share. This is an important value because many intrinsic aspects of the product are coded within the past values. For instance, its quality, its notoriety, its position on the market are reflected in the past sales and they do not change rapidly. In practice, we would consider more than 1 value in the past, which mean that we would have : $w_{i,t} = \phi(\widetilde{y_{i,t-k}}, \dots, \widetilde{y_{i,t-1}}, \theta_{i,t})$.

The second parameter is the vector of covariates $(\theta_{i,t})$. These covariates should explain the variation of competitiveness.

Finally, let's remark that we consider that the underlying behavior of each series is the same. Indeed, we use a unique function ϕ for all the series we observe instead of using a specific ϕ_i for every series. It means that the different series are interchangeable and have the same behavior. It allows them to share information between series and to adapt to newly introduce the product on the market.

However, this last assumption has some drawbacks. In particular, when some unknown or unrecorded features are relevant for the prediction of the sales, this could lead to changes in the market share that are not fully explained by this model.

6.2.5 Summary

The model may thus be written as :

$$\begin{cases} X_1 & \sim \mathscr{P}_{X_1} \\ x_{i,t} & = \epsilon_{i,t}(\lambda_{i,t}) \\ \lambda_{i,t} & = s(t)\dfrac{\phi(\widetilde{y_{i,t-1}}, \theta_{i,t})}{1 + \sum\limits_{j\in[1,d]} \phi(\widetilde{y_{i,t-1}}, \theta_{j,t})} \end{cases} \quad \text{for } t > 1$$

where $\mathscr{P}_{X_1}$ denotes the distribution of the first values. With this model (X_t) is a (non-homogeneous) Markov chain with a transition function F_t such that

$$X_t = F_t(X_{t-1}, \epsilon_t)$$

where $\epsilon_t = (\epsilon_{1,t}, \dots \epsilon_{i,t})$. More precisely

$$F_t(X, \epsilon_t) = \epsilon_t \left(s(t)\frac{\Phi(\frac{X}{s(t-1)}, \theta_t)}{1 + \|\Phi(\frac{X}{s(t-1)}, \theta_t)\|} \right) \tag{6.2}$$

where Φ is the extension of ϕ to $\mathbb{R}_+^d$.

Comparison with other models This model extends both univariate count model with exogeneous covariates [ACKR16; PR19; DNT20] and non-linear univariate count models [Fok12]. It is also an extension of multivariate count auto-regressive model [FSTD20], where we add non-linear relation between times series.

We also paralleled this model with the model proposed for auto-regressive categorical times series introduced by [FT17]. They extend this model using covariates, prooving the ergodicity and stationarity of Markovian model with covariates under mixing conditions. However, the main difference is the use of a logistic function instead of a simple ratio between weights. Logistic regression often appears in classification, because it tends to assign all the weight to a single class. Here we want to model an equilibrium between different classes and therefore we do not use a logistic function.

6.3 Estimation Risk bounds on Empirical Risk Estimator

In this section, we establish theoretical bounds on the estimation risk of our model. Contrary to [DFT12], we cannot use weak dependence hypotheses. As in [ACKR16; DT20], we will use introduce a contraction condition to control the dispersion of the model. The specific contraction and exponential inequalities we will use were introduced by Dedecker and Fan in [DF15] and extended for the non-stationary times series in [ADF19].

6.3.1 Contraction condition

In order to apply the result of [ADF19], a contraction condition on the Markov transition function must be verified. More precisely, there must be a constant $\rho \in [0,1[$ such that for all $X, X' \in \mathbb{R}^d$:

$$\sup_t \mathbb{E}\Big[\|F_t(X,\varepsilon_t) - F_t(X',\varepsilon_t)\|\Big] \leq \rho \|X - X'\| \tag{6.3}$$

This condition holds under additional assumptions over the function ϕ

Lemme 6.1. *Assume that we have a weight function ϕ and a seasonality s defining a transition function F_t as (6.2). If ϕ and s are such that :*

1. *There is a constant τ_s such that, for all t :*

$$\frac{s(t+1)}{s(t)} \leq \tau_s$$

,

2. *There is a constant $\tau \in \mathbb{R}_+$ such that for all $x \in \mathbb{R}_+, \theta \in \mathbb{R}^p$*

$$\frac{\delta \phi}{\delta x}(x,\theta) \leq \tau$$

,

3. *For all $X \in \mathbb{R}_+^d$ and $\theta \in \mathbb{R}^p$, $\|\Phi(X,\theta)\| \geq 1$*

4. $\tau_s \tau < 1$.

Then the random iterated system F_t fits the contraction condition (6.3) for any $\rho < 1$ such that $\tau_s \tau \leq \rho$.

Démonstration. Note $G_t(X) = \frac{\Phi(\frac{X}{s(t-1)}, \theta_{i,t})}{1+\|\Phi(\frac{X}{s(t-1)}, \theta_i)\|}$

For $X, X' \in {\mathbb{R}_+}^d$

$$
\begin{aligned}
\mathbb{E}\Big[\|F_t(X,\varepsilon_t)-F_t(X',\varepsilon_t)\|\Big] \leq\ & s(t)\|G_t(X)-G_t(X')\| \\
=\ & s(t)\frac{\|(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)\Phi(\frac{X}{s(t-1)},\theta)-(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)\Phi(\frac{X'}{s(t-1)},\theta)\|}{(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)\cdot(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)} \\
\leq\ & s(t)\frac{\|(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)\Phi(\frac{X}{s(t-1)},\theta)-(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)\Phi(\frac{X'}{s(t-1)},\theta)\|}{(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)\cdot(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)} \\
& + s(t)\frac{\|(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)\Phi(\frac{X'}{s(t-1)},\theta)-(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)\Phi(\frac{X'}{s(t-1)},\theta)\|}{(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)\cdot(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)} \\
\leq\ & s(t)\frac{\|\Phi(\frac{X}{s(t-1)},\theta)-\Phi(\frac{X'}{s(t-1)},\theta)\|}{(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)} \\
& + s(t)\frac{\|\Phi(\frac{X'}{s(t-1)},\theta)\|(\|\Phi(\frac{X'}{s(t-1)},\theta)\|-\|\Phi(\frac{X}{s(t-1)},\theta)\|)}{(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)\cdot(1+\|\Phi(\frac{X'}{s(t-1)},\theta)\|)} \\
\leq\ & 2s(t)\frac{\|\Phi(\frac{X}{s(t-1)},\theta)-\Phi(\frac{X'}{s(t-1)},\theta)\|}{(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)}
\end{aligned}
$$

Using the condition 2, $x \mapsto \phi(x,\theta)$ is τ-Lipschitz for each θ.

$$
\mathbb{E}\Big[\|F_t(X,\varepsilon_t)-F_t(X',\varepsilon_t)\|\Big] \leq \frac{2}{(1+\|\Phi(\frac{X}{s(t-1)},\theta)\|)}\frac{s(t)}{s(t-1)}\tau\|X-X'\|
$$

Using the above conditions 1, 3 and 4, we have :

$$
\mathbb{E}\Big[\|F_t(X,\varepsilon_t)-F_t(X',\varepsilon_t)\|\Big] \leq \tau\tau_s\|X-X'\| \leq \rho\|X-X'\|
$$

which concludes the proof. □

We now discuss the conditions in Lemma 6.1.
The first set a restriction on the regularity of the seasonality, which should not change too abruptly. In practice, this is not always verified. Indeed, some event like Black Friday creates drastic changes in product sales seasonality. Ideally, such event should be handled with other techniques. Otherwise, changes in seasonality are mostly smooth.
The second and the fourth condition limit the variations of the weight function. This is a constraint on the type of models that we use to build this weight function, since it should be smooth. In particular, tree-based models such do not fit this condition. However, for neural network this is similar to the stability condition described in [MH19], where the author show that enforcing this conditions does not degrade the performance of recurrent neural network.
The third condition is a direct consequence of the '+1' trick we use for stabilizing the weight. It is necessary to ensure that this '+1' does not suppress the rest of the weights. In theory, this condition is very strong, because it implies a strong constraint on the whole domain of the function Φ to $\mathbb{R}_+^d$. However, in practice, the domain of application of this function is more restrained, and we just need to have this condition on this restrained set.

To compute a generalisation bound on our model we need to introduce :

$$
\begin{aligned}
G_{X_1}(x) &= \int \|x-x'\| dP_{X_1}(dx') \\
H_{t,\varepsilon_t}(x,y) &= \int \|F_t(x,y)-F_t(x,y')\| dP_{\varepsilon_t}(dy')
\end{aligned}
$$

To have a Bernstein inequality, we need some constraints on the dispersion of our the times series. More precisely, we need to have the following inequalities for some constants $M > 0$, $V_1 > 0$ and $V_2 > 0$ such that, for all integer $k \geq 2$:

$$\mathbb{E}[G_{X_1}(X_1)^k] \leq \frac{k!}{2} V_1 M^{k-2} \tag{6.4}$$

$$\forall t, \forall x, \mathbb{E}[H_{t,\varepsilon_t}(x,\varepsilon_t)^k] \leq \frac{k!}{2} V_2 M^{k-2} \tag{6.5}$$

The condition (6.4) is satisfied if X_1 admits subgaussian moments. However, this condition is less restrictive than sub-Gaussianity. Now let us consider the second condition (6.5).

Lemme 6.2. *If there exists* $R > 0$ *such that, for each* t, $s(t) \leq R$, *and if we denote by* $M = d \max\{1, eR\}$ *and* $V_2 = 4M^2 = 4d^2 \max\{1, eR\}^2$, *then :*

$$\mathbb{E}[H_{t,\varepsilon_t}(x,\varepsilon_t)^k] \leq \frac{k!}{2} V_2 M^{k-2}$$

Démonstration. Let us consider

$$\mathbb{E}[H_{t,\varepsilon_t}(x,\varepsilon_t)^k] = \int \left(\int \|F_t(x,y) - F_t(x,y')\| dP_{\varepsilon_t}(dy') \right)^k dP_{\varepsilon_t}(dy)$$

From Jensen inequality we derive :

$$\begin{aligned} \mathbb{E}[H_{t,\varepsilon_t}(x,\varepsilon_t)^k] &\leq \int \int \|y(s(t)G_t(x)) - y'(s(t)G_t(x))\|^k dP_{\varepsilon_t}(dy') dP_{\varepsilon_t}(dy) \\ &\leq \mathbb{E}[\|Y - Y'\|^k] \end{aligned}$$

where $Y = (Y_i)$ and $Y' = (Y_i')$ are independent vector of independent random variables following Poisson distributions with parameters $s(t)G_t(x)$. Hence :

$$\mathbb{E}[H_{t,\varepsilon_t}(x,\varepsilon_t)^k] \leq \mathbb{E}[\|Y\|^k + \|Y'\|^k] \leq 2\mathbb{E}[\|Y\|^k].$$

As $\|G_t(X)\|_\infty \leq 1$, we have

$$\mathbb{E}[\|Y\|^k] \leq d^k \mathbb{E}[\|Y\|_\infty^k] \leq d^k \mathbb{E}[(y_t)^k]$$

where y_t is a random variable following a Poisson process with parameter $s(t)$.

Using Lemma C.1, we obtain $\mathbb{E}[(y_t)^k] \leq k! \max\{es(t), 1\}^k$. This ensures

$$\mathbb{E}[H_{t,\varepsilon_t}(x,\varepsilon_t)^k] \leq 2d^k k! \max\{es(t), 1\}^k,$$

and this allows us to conclude. □

Thanks to this lemma, we just need the seasonality $s(t)$ to be bounded to have the condition (6.5). Nevertheless, we must note that we use loose bounds. In particular, V_2 may be great when the dimension of the serie d is important. Getting a better value for V_2 would be necessary to get a tighter risk bound.

6.3.2 Risk Bounds on Empirical Risk Estimator

In this section, a bound on model selection error is provided.

Let (X_t) be an $\mathbb{R}_+^d$ valued process with n observations following the model described in part 1 for a function ϕ^*. Let S be a set of functions respecting the condition of the Lemma 1 such that $\phi^* \in S$. For a function $\phi \in S$, we define an empirical risk :

$$R_n(\phi) = \frac{1}{n}\sum_{t=2}^{n} \|X_{t+1} - s(t+1)\frac{\Phi(\frac{X_t}{s(t)}, \theta_t)}{1 + \|\Phi(\frac{X_t}{s(t)}, \theta_t)\|}\|.$$

We also define :

$$R(\phi) = \mathbb{E}[R_n(\phi)]$$

We define the minimum empirical risk estimator :

$$\widehat{\phi} = \underset{\phi \in S}{\operatorname{argmin}}\, R_n(\phi)$$

It is possible to bound the estimation risk :

Théorème 6.1. *Let* $K_t(\rho) = \frac{1-\rho^t}{1-\rho}$. *If* ϕ *and* $X = (X_i)$ *verified the condition (1) to (4), then for* $\delta > 0$ *we have with probability* $1-\delta$ *:*

$$R(\widehat{\phi}) \leq R(\phi^*) + (1+\tau)\left(\frac{\sqrt{2V_2 \log(\frac{1}{\delta})}}{\sqrt{n}} + \frac{\sqrt{2V_1 \log(\frac{1}{\delta})}}{n} + \frac{2MK_{n-1}(\rho)\log(\frac{1}{\delta})}{n}\right)$$

Remarque 6.1. *We observe the usual decay in* $\mathcal{O}(\sqrt{\frac{\log(\frac{1}{\delta})}{n}})$. *If we use the values for* V_1 *established in the Lemma 2, we observe that the error grows linearly with the dimension* d.

Démonstration. First, let's recall the usual argument to bound the excess risk :

$$\begin{aligned} R(\widehat{\phi}) - R(\phi^*) &= R(\widehat{\phi}) - R_n(\widehat{\phi}) + R_n(\widehat{\phi}) - R_n(\phi^*) + R_n(\phi^*) - R(\phi^*) \\ &\leq |R_n(\phi^*) - R(\phi^*)| + |R(\widehat{\phi}) - R_n(\widehat{\phi})| \qquad \text{(from the definition of } \widehat{\phi}) \end{aligned}$$

Therefore for all $t > 0$, it holds :

$$\begin{aligned} \mathbb{P}[R(\widehat{\phi}) - R(\phi^*) \geq t] &\leq \mathbb{P}[|R_n(\phi^*) - R(\phi^*)| + |R(\widehat{\phi}) - R_n(\widehat{\phi})| \geq t] \\ &\leq \mathbb{P}[|R_n(\phi^*) - R(\phi^*)| \geq \frac{t}{2}] + \mathbb{P}[|R(\widehat{\phi}) - R_n(\widehat{\phi})| \geq \frac{t}{2}] \end{aligned}$$

Thus :

$$\mathbb{P}[R(\widehat{\phi}) - R(\phi^*) \geq t] \leq 2 \sup_{\phi \in S} \mathbb{P}[|R(\phi) - R_n(\phi)| \geq \frac{t}{2}] \tag{6.6}$$

Then, we aim at bounding the difference $R_n(\phi) - R(\phi)$ for all possible functions ϕ.

Let $(\mathcal{F}_k)$ be the natural filtration of the chain (X_k)

R_n is $\frac{(1+\tau)}{n}$ Lipschitz separable. Therefore for $\epsilon > 0$, we can apply the theorem 3.1 of [ADF19] to $\frac{R_n}{(1+\tau)}$. Actually, we use a slightly different version, as the space of Poisson processes are not actually separable. However, being able to bound $\mathbb{E}[H_{t,\varepsilon_t}(x,y)^k]$ suffice to use their version of Bernstein inequality.

$$\mathbb{P}[|R_n(\phi) - R(\phi)| \geq \frac{(1+\tau)}{n}\epsilon] \leq \exp\left(\frac{-\epsilon^2}{2V_1 + 2(n-1)V_2 + \epsilon MK_{n-1}(\rho)}\right)$$

Hence, we have with probability at least $1-\delta$:

$$|R_n(\phi) - R(\phi)| \leq \frac{(1+\tau)\sqrt{(2V_1 + 2(n-1)V_2)\log(\frac{1}{\delta})}}{2n} + \frac{(1+\tau)MK_{n-1}(\rho)\log(\frac{1}{\delta})}{n}$$

Therefore, using (6.6), with probability at $1-\delta$, we have :

$$R(\widehat{\phi}) \leq R(\phi^*) + (1+\tau)\left(\frac{\sqrt{2V_2\log(\frac{2}{\delta})}}{\sqrt{n}} + \frac{\sqrt{2V_1\log(\frac{2}{\delta})}}{n} + \frac{2MK_{n-1}(\rho)\log(\frac{2}{\delta})}{n}\right)$$

□

6.4 Implementation of the model

Now we present the implementation of the theoretical model proposed in section 6.2 and we show how it could be used for times series prediction. The code is available on Github [Git].

6.4.1 Empirical risk minimization

We want to adapt our model to a classical machine learning setting, using empirical risk minimization, in order to be able to use an efficient optimization algorithm. To do so, we introduce a set Ψ of possible weight function ϕ. We will search for the optimal function in this set.

We will perform a prediction at an horizon $h \geq 1$. For a chosen weight function ϕ a covariate vector $\theta_{i,t}$ and known past value of the market share $y_{i,t-h}$, the next value will be predicted by :

$$\hat{y}_{\phi,i,t} = \frac{\phi(y_{i,t-h}, \theta_{i,t})}{1 + \sum_{j\in[1,d]} \phi(y_{j,t-h}, \theta_{j,t})}$$

Contrary to the previous section, we use actual market share instead of approached values $\widetilde{y_{i,t}}$. It doesn't seem to change the results in any meaningful way. We will then perform the empirical risk minimization for a loss L :

$$\hat{\phi} = \underset{\phi\in\Psi}{\operatorname{argmin}} \sum_{t=1}^{n-h}\sum_{i=1}^{d} L(y_{i,t}, \hat{y}_{\phi,i,t}) \tag{6.7}$$

Note that, when L is the Poisson loss function $L(y,\hat{y}) = \hat{y} - x\log\hat{y}$, the empirical risk minimizer is also the function ϕ which minimize the log likelihood of the model presented in section 6.2. However, we will also use the more standard L_1-loss function to show the interest of the Poisson distribution.

6.4.2 Concurrent Neural network

The most complex choice is that of the set Ψ among which we choose the weight function ϕ. It should be able to satisfy several properties.

Firstly, it should be complex enough in order to handle non linear behavior. Indeed, we want to model complex behavior, that depends interacting and sometimes correlated parameters.

Secondly, the considered functions must be differentiable. This is a condition necessary to use powerful optimization algorithm. This condition disqualifies most of tree-based models, often used to predict sales in different context [Van21].

This constraint leads us to use feed-forward neural network as sub-models ϕ. Around them we build a structure that we called a concurrent neural network model that we will note Conc-NN to distinguish it from Convolution Neural Network traditionally abbreviated CNN. We summarize this approach on figure 6.1.

We introduce a scale factor $\alpha < 1$. This is because the sums $\sum_{i=1}^{d} y_{i,t-h}$ may be smaller than 1 in some case from of the presence of newly introduced product between the date where the prediction is made $t-h$ and the date where the prediction is actualized t.

In order to stay simple and not introduce any bias in the comparison between models, we do not choose a data-driven values for α. For short and medium-term horizon, $\alpha = 1$, but for long term horizon, we consider $\alpha = 0.85$.

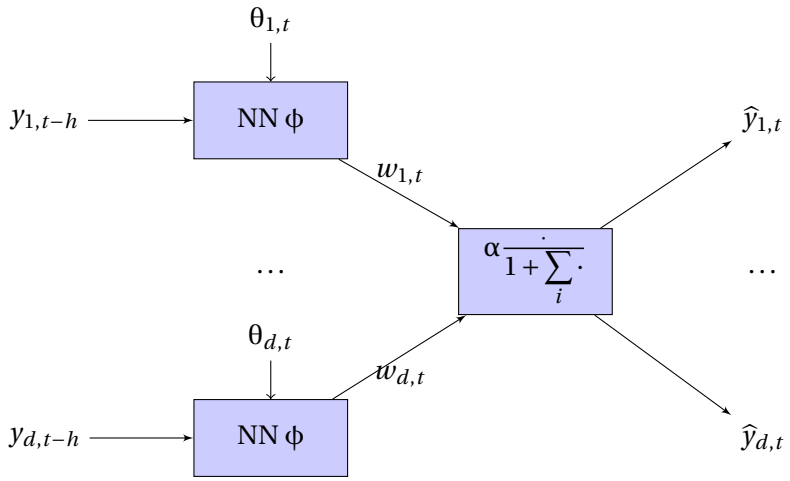

FIGURE 6.1 – Concurrent NN Model

6.4.3 Neural network architecture and training

Four neural network estimators are presented in the results.

— **FF-NN** Classical Feed-forward neural network trained with L1 Loss

— **L1-Conc-NN** Concurrent neural network presented in the last subsection trained with L1-Loss

— **P-Conc-NN** Same model trained with Poisson Loss

— **L1-Pre-Conc-NN** Concurrent neural network trained with L1-Loss using the pre-trained weight of FF-NN

The training methods are the same for all those models. We use simple feed-forward architecture with less than four layers, and less than 32 neurons per layers. We use a RELU activation function for every layer except the last one where we use a SoftPlus activation. For every category of products, we use a validation period to perform model selection among 10 different neural network architectures. Data are introduced by batch in a random order. Every batch correspond to a week, to allows an easy rescaling in the case of Concurrent Neural Network.

6.5 Application to E-Commerce sales dataset

In this section, we will try to apply our method to a real dataset of E-commerce sales. First, we present the data used in our application.

6.5.1 Datasets

We consider different data sets coming from the E-commerce company *Cdiscount*. It is a French E-commerce retailer selling a large variety of products.

We use the data available for different families of products sold by CDiscount. These categories have been selected to represent various types of products. The hyper-parameters of the models were chosen using other families. The product categories are presented in Table 6.1, along with the number of products d and some descriptive statistics. SD stands for Standard Deviation, where NSD (Normalized Standard Deviation) is the ratio between SD and the Weekly sales average by product. We also comupute DS-NSD (Deseasonalized Normalized Standard Deviation) the same ratio for deseasonalized weekly sales. Average and standard deviation are computed only when the products are actually proposed on the website.

These datasets can be roughly separated into three categories. The first one is the product that presents regular seasonality and where the demand is relatively insensitive to price variation. It is

Family	d	Total weekly sales average	Weekly sales average by product	Max sales	SD	NSD	DS-NSD
Baby chair	127	483.0	9.8	238	16.3	1.66	1.50
Freezer	139	1364.0	23.9	567	39.3	1.64	1.45
Keyboard	68	267.8	9.3	374	19.2	2.06	1.55
Lawn Mower	81	369.3	13.5	455	27.4	2.02	1.30
Scooter	589	1927.1	8.8	785	20.3	2.30	1.93
SD Cars	45	288.6	16.4	542	36.1	2.2	1.64
Smartphone	1055	8352.3	29.8	2886	81.6	2.74	2.37
TVs	535	8004.9	64.8	5547	148.2	2.29	1.95

TABLEAU 6.1 – Datasets descriptive statistics

the case for Baby chairs and Freezers, which have a small NSD and DS-NSD, and where NSD and DS-NSD are relative similar. The second one is products that present strong seasonality factors, such as Lawn Mowers and Scooters, but are not highly sensitive to price changes. In this case, we observe an important difference between NSD and DS-NSD due to the seasonality factor. The last ones are products such as Smartphones and TVs, which present short sales cycles and/or are very sensitive to price changes, which translates into important NSD and DS-NSD.

We consider the weekly sales starting from January 2017 to December 2020. The first three years are used to train the models, which are evaluated on the last year of data. Added external features include the margin practiced on the product and their prices.

6.5.2 Evaluation

We want to predict the weekly sales shares of different products for an horizon of h weeks. We will evaluate the prediction using the usual Mean Absolute Error(%MAE) . For a prediction $\widehat{y}_{i,t+h}$, it is defined as :

$$\%\text{MAE} = 100 * \frac{\sum_{i=0}^{d}\sum_{t=0}^{\text{T}-h}|\widehat{y}_{i,t} - y_{i,t}|}{\sum_{i=0}^{d}\sum_{t=0}^{\text{T}-h} y_{i,t}}$$

The %MAE has two main advantages. First, it can deals with outliers, which tends to be overweighted with other metrics such as RMSE and NMRSE. First, it scales with the level of sales, and so can be used to compare the prediction for the different product categories. However, %MAE error tends to favor under-estimated predictions.

We will use other predictors to get a benchmark of prediction.

— **Last Value (LV) :** Use the last known value to predict future market share.

— **Moving average (MA) :** Use a moving average model to predict the future values. Hyperparameters are calibrated on a validation period.

— **Random forest (RF) :** Random forest using the same features as ConcNN. Different architectures are cross-validated on a validation period.

— **Scaled Random forest (S-RF) :** We also use Random Forest where the prediction are scaled to match the total number of sales. This is useful to compare our models with the results when we perform a simplistic re-scaling after a model is trained.

Category	LV	MA	RF	S-RF	FF-NN	L1-Conc-NN	P-Conc-NN	L1-Pre-Conc-NN
Baby chair	76.7	73.8	70.7	70.5	67.7	**59.9**	63.5	*
Freezer	88.1	85.3	69.7	70.1	67.1	66.3	**65.2**	68.3
Keyboard	87.6	81.7	76.7	77.8	**67.1**	72.7	70.3	*
Lawn Mower	83.2	81.5	72.8	75.7	74.6	75.6	**72.5**	74.1
Scooter	86.3	84.1	78.5	82.0	75.2	**73.6**	75.7	74.0
SD Cars	83.7	79.3	88.9	90.3	**74.0**	75.8	77.9	*
Smartphone	84.0	81.6	79.1	84.3	81.3	**75.6**	75.8	75.7
TVs	78.5	80.7	76.2	78.2	*	**71.7**	77.8	*

TABLEAU 6.2 – %MAE Results on the market share prediction for an short-term horizon $h = 4$ weeks

Category	LV	MA	RF	S-RF	FF-NN	L1-Conc-NN	P-Conc-NN	L1-Pre-Conc-NN
Baby chair	85.03	82.4	79.8	80.4	81.2	78.3	**75.6**	*
Freezer	97.5	94.6	76.7	80.2	**70.6**	71.4	71.0	72.3
Keyboard	82.9	79.5	76.5	76.6	**68.0**	70.1	75.3	70.6
Lawn Mower	96.0	92.3	**76.8**	83.0	*	83.0	82.1	*
Scooter	99.5	97.1	79.6	86.7	**76.5**	77.8	76.9	77.9
SD Cars	89.1	86.7	89.5	89.9	79.0	**78.3**	79.0	*
Smartphone	93.6	90.1	88.2	95.6	88.3	85.4	85.3	**84.5**
TVs	103.5	104.0	89.9	97.4	*	**81.4**	90.8	*

TABLEAU 6.3 – %MAE Results on the market share prediction for an medium-term horizon $h = 8$ weeks

6.5.3 Results

We present the results in the Table 6.2 for the horizon $h = 4$ (1 month ahead),in the Table 6.3 for horizon $h = 8$ (2 months ahead), and in the Table 6.4 for the horizon $h = 12$ (3 months ahead). In some cases, FF-NN were not able to produce any meaningful prevision and only predict 0. In this case, we put a star (*) in the columns. The best model for every set of products and every horizon is in **bold**.

L1-Pre-Conc-NN also failed to produce any prevision other than 0. It is of course the case when FF-NN provides zero predictions, but it can also happen when pretrained weights are too small. We also denote this case by a (*).

General Remarks. As expected, the error increases with the horizon of prediction h. It is expected, as long-term previsions are generally more complex than short-term previsions. However, let us remark that this increasing complexity is not the same for every category. When sales cycles are short, for instance for TVs, the influx of new products makes long-term prediction even harder.

Comparison with LV, MA The different Conc-NN models outperform both classical times series estimators LV and MA for almost every horizon and products sets. In particular, traditional time series estimators. It means that Conc-NN can exploit external features.

Category	LV	MA	RF	S-RF	FF-NN	L1-Conc-NN	P-Conc-NN	L1-Pre-Conc-NN
Baby chair	88.9	86.0	78.2	77.8	77.1	**76.1**	80.0	85.9
Freezer	97.7	94.7	73.1	76.0	72.3	70.9	**70.0**	78.1
Keyboard	91.5	87.0	79.5	79.1	**69.4**	69.9	74.8	*
Lawn Mower	99.3	96.9	**80.2**	86.4	82.7	84.5	81.5	84.5
Scooter	99.5	97.1	**79.8**	87.1	80.5	85.4	84.0	83.8
SD Cars	95.3	91.6	86.4	89.4	82.2	**77.2**	79.4	*
Smartphone	97.9	95.0	91.5	100.9	89.7	85.6	**85.3**	90.5
TVs	117.4	116.8	**97.5**	115.1	*	101.8	110.6	*

TABLEAU 6.4 – %MAE Results on the market share prediction for an long-term horizon $h = 12$ weeks

Comparison with RF The different Conc-NN models outperform random forests (RF) for almost every product for short-term prediction, but RF becomes better for longer-term horizon. One way to explain this fact is that RF tends to under-predict sales, which favors it for %MAE evaluation. Example of such under-prediction are presented in figure 6.2

Note that S-RF prediction strongly under-performs RF. Therefore, simply rescaling prediction is not enough to correctly distribute market share.

Comparison with FF-NN. For smartphones and TVs, Conc-NN outperforms FF-NN. These categories are considered as the most competitive, with a lot of price changes and short sales cycles. It may be the proof that our model correctly describes the competition mechanisms in this category.

There is maybe another explanation, however. Conc-NN also performs well for Scooters, and the three categories (Scooter, TVs, and Smartphones) are also the categories with the most products. They also have good performances on SD Cards, which is also a very competitive category with a few products.

FF-NN outperforms Conc-NN models for keyboards for every horizon.

FF-NN shares a drawback with RF. It under-predicts some products. This may be also due to the L_1-Loss minimization, which tends to favor under-predictive models. We also show some examples in Figure 6.2.

Pretrained Model L1-Pre-Conc-NN. Pretrained Concurrent model generally under-performs other Conc-NN. Most of the time, they also under-perform the FF-NN used for pretraining. They also tend to predict 0 a lot.
Therefore, pretraining models do not seem to be generally useful. However, they obtain generally good results for Smartphones for all horizons.

Poisson Loss VS L1 Loss. Poisson and L1 Loss leads generally to similar performances. It is hard to see any pattern in the relative performance of both models. Let us just notice that L1 Loss outperforms Poisson for every horizon for TVs, whereas Poisson Loss outperforms L1 for Lawn Mowers, but it could be explained by accident.

When we observe the prediction in Figure 6.2, we can notice that L1-Loss-based prediction tends to present higher variation than Poisson-Loss-based prediction. This sensitivity to variation may explain the higher performance of L1-Loss for TVs.

6.5.4 Features Importance

Our models take into account two external features, the price and margin practiced on the product. To understand how our model treats the covariate, we consider the partial dependence of our models to see if we could explain the variation of underlying weight function ϕ. To compute the partial dependance graph for a feature θ, we use the following procedures :

— We consider the distribution of a given feature θ in the training set, and we split this distribution into 100 bins.

— For each bin i we compute the average value θ_i of the feature θ on the bins.

— We compute the average weight on the test set of $\phi(\tilde{\theta}_{i,t})$, where $\tilde{\theta}_{i,t}$ is the usual point of data of the test set where the feature θ has been replaced by θ_i

— We then plot all the couple (average bin , average weight) .

Note that when we compute the partial dependance with respect to price and margin, the function ϕ is computed without using the other features. We do so to avoid the effect of co-integration of the variables. All the partial dependence are computed for two categories : smartphones and large freezers, for an horizon $h = 4$.

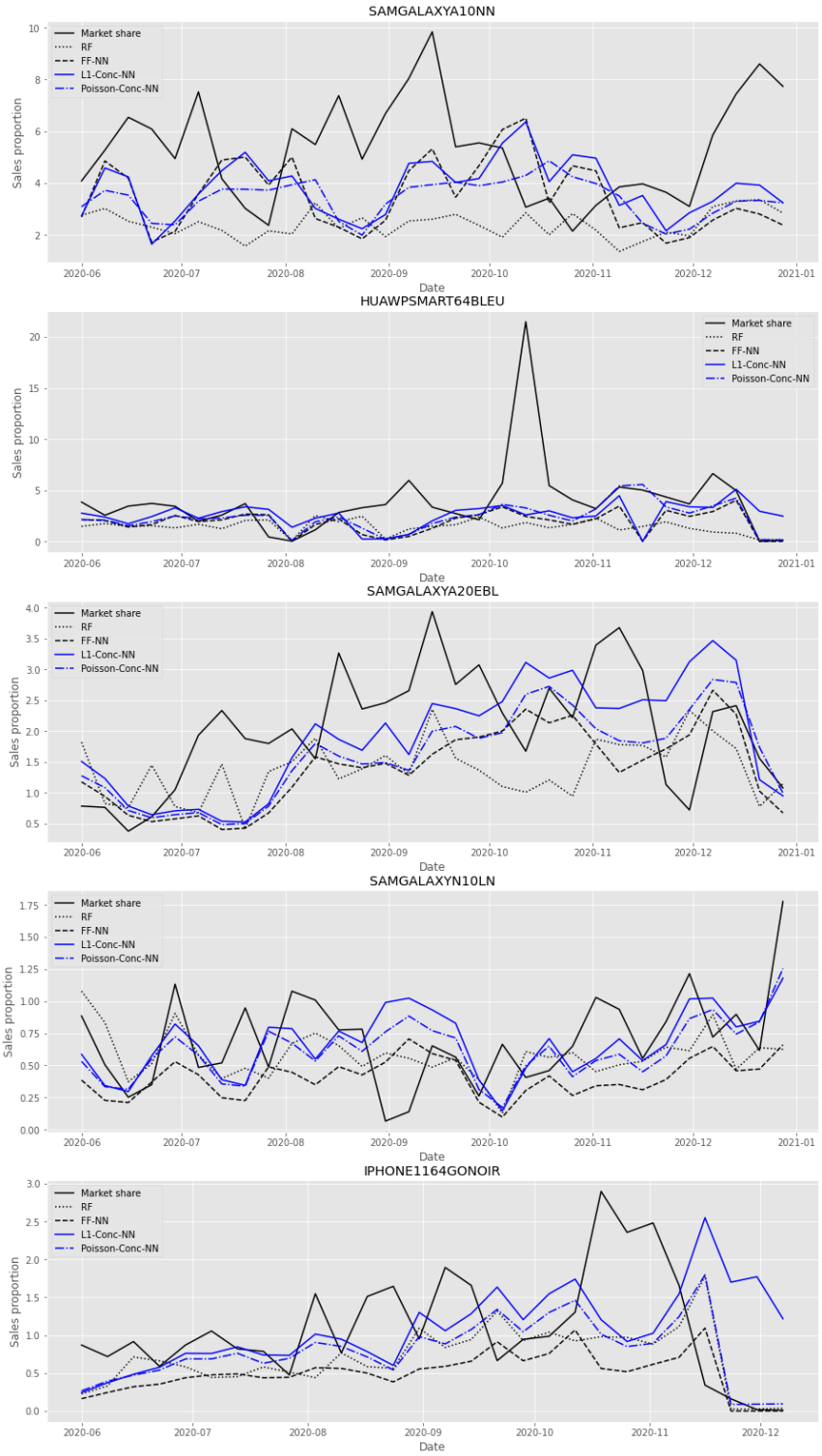

FIGURE 6.2 – Market prediction for some popular smartphones for an horizon $h = 4$

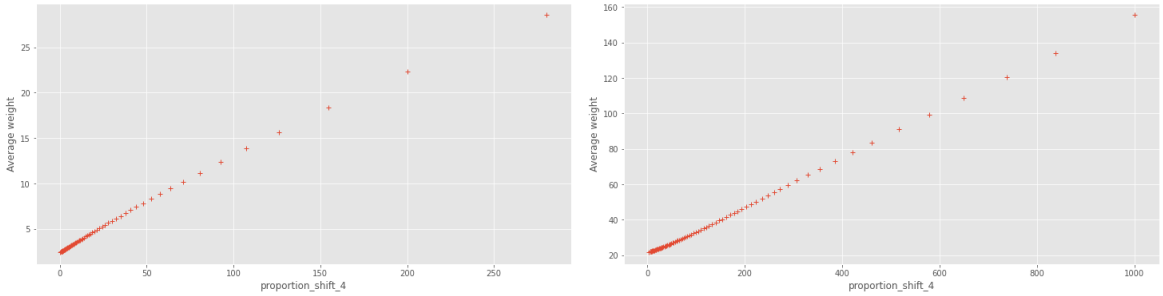

FIGURE 6.3 – Partial dependence with respect to the past proportion variables (on the left for phones, on the right for large freezers)

Partial dependence w.r.t. past proportion. On Figure 6.3, we present the partial dependence of the weight with respect to the last known market share. Logically, it is increasing. When the past sales proportion of the sales were important, it is a strong indication of the competitiveness of a product.

Partial dependence w.r.t. prices. On Figure 6.4, we present the partial dependence of the weight with respect to the price of the smartphone.

Smartphones reach their peak competitivity around 150 €, then the average weight decrease. Lower prices seem to indicate a very poor quality of the smartphones, therefore reducing the attractivity for customers. Logically, higher prices also decrease attractivity.

This is similar for freezers, with a peak attractivity around also 150 €. However, there is small increase in competitivity with very high price. It could be a Veblen effect, or an effect of the catalog composition.

Partial dependence w.r.t. margin. On Figure 6.5, we present the partial dependence of the weight with respect to the margin of the smartphone. Margin is a better indicator than prices, because it is insensitive to the relative quality of the products [1].

For smartphones, when margin is positive, the weight seems to increase with the margin. It is also counter-intuitive, but can also be explained by pricing behavior of the company. When a product receive a great interest, the company can easily increase its margin. The company may also want to push forward products with higher margin.

When margin are negative however, i.e. when products are on sales, we observe a strong increase on its competitiveness. This is expected, as negative margins means that the product are on sales.

For freezers, the behavior is closer to expected behavior when margin is higher than 10%. However, for low margin, we observe a strong decrease in competitivity. It may also be explained by the pricing strategy.

1. For an E-Commerce website, the relative marge are not higher on high quality products. This is not the case for a manufacturer, who tends to make higher margin on premium products.

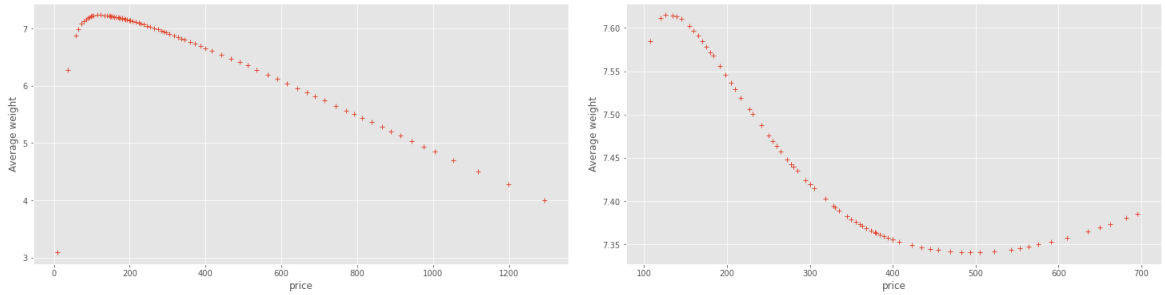

FIGURE 6.4 – Partial dependence with respect to the prices (on the left for phones, on the right for large freezers)

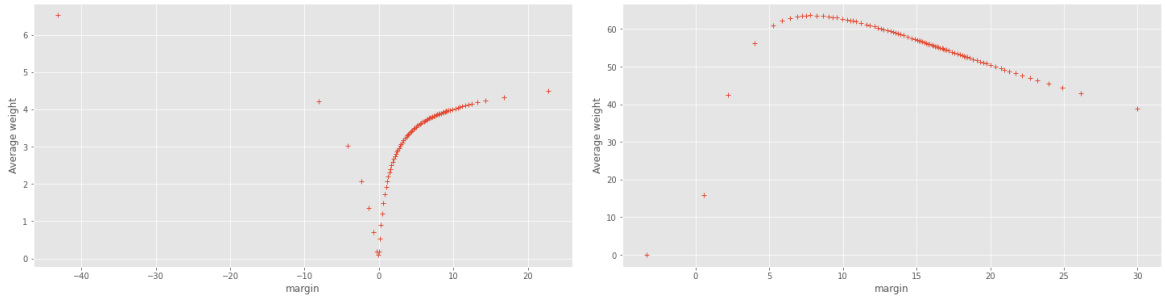

FIGURE 6.5 – Partial dependence with respect to the margin (on the left for phones, on the right for large freezers)

6.6 Conclusions

A model for concurrent time series is proposed in this article. It is based on the creation of unknown "competitiveness" quantity that depends on the characteristic of a product and previous information on its popularity. Under a relatively common condition, we establish a bound on the risk of our model. This bound follows the usual decay in $\mathcal{O}(\sqrt{\frac{\log(\frac{1}{\delta})}{n}})$ observed in Machine Learning.

We use this model on real-world data, using a Neural Network to compute the competitiveness using past sales values and other external features. This approach outperforms classical ML and times series estimators, especially for short and medium terms predictions. It also improves classical Neural Network estimators, especially when they are numerous products and when the competition between time series is high. It also partially avoid under-prediction that tends to affect other predictors.It is possible to explain the weight function. However, the behavior of this function may be counter-intuitive due to the feedback loop.

Further work may include the extension of this model to non-differentiable sub-model such as Random Forests or Boosting trees. Indeed, practitioners often use a mix of categorical and numerical features to predict sales, and this could only be done by a non-differentiable model. Therefore, extending concurrent model could be of great use for practionners.

Chapitre 7

Perspectives

Il est compliqué de résumer tout les travaux de la thèse qui peuvent paraître bien disparates et concernent des domaines finalement assez différents. On va donc essayer de dégager des perspectives variés à partir de chacune des contributions, ou de combinaison de plusieurs de ces contributions.

Graphes non-causaux et prévision des ventes par graphe Commençons par quelque chose qui n'a pas pu être fait, et qui explique pourquoi un chapitre sur des données non-causales se retrouve dans une thèse sur la prévision des ventes. Une des idées qui dirigeait l'utilisation de modèles non-causaux était l'idée qu'on pouvait considérer les ventes d'un ensemble de produit à une date t donnée comme une structure présentant des relations non-causales. Ainsi, on peut modéliser les ventes avec des variables aléatoires placées dans une structure de graphe fixes, chaque produit étant un noeud du graphe [1]. Le cadre n'est pas tout à fait analogue à ce qui a été présenté chapitre 3, parce que notre ensemble de produit n'est pas muni d'une relation d'ordre [2] ou placé sur un treillis. Cependant, l'extension à des graphes quelconque de ce qui a été fait à ce chapitre pour des treillis ne parait pas impossible, à condition que le degré des noeuds du graphe reste borné. Il semble possible d'appliquer l'idée de l'approximation sur des sous-graphes concentriques à distance de plus en plus grandes du noeud central.

Cependant, il y a deux problèmes pour appliquer cette voie à la prévision des ventes :

— D'abord, la variabilité du graphe. Du fait de l'entrée et la sortie continue de nombreux produits du catalogue, la structure du graphe sous-jacent est extrêmement mobile.

— Ensuite parce que l'hypothèse de stationnarité sur les noeuds du graphe ne tient pas. Certains produits ont nettement plus de ventes que d'autres. Le degré variable des noeuds est une autre raison qui oblige à travailler avec des données non stationnaires.

En fait, les hypothèses qui rendent le travail théorique du chapitre 3 possible rendent compliqués à appliquer à des prévisions de la demande. Il aurait donc été compliqué d'unifier en unique contribution le chapitre 3 avec une approche "par graphe" de la prévision.

Par ailleurs, l'approche graphe est une approche qui modélise les interactions des différents paires produits . Cette type de modelisation d'interaction a certainement un intérêt, puisqu'elle est à la base du travail de [BLP95] dans le secteur automobile. Cependant, la quantité de données disponible et surtout le nombre de produits importants rendait peu praticable une telle approche chez Cdiscount. Pour cette raison, on a été amené à ne pas travailler comme cela pour modéliser les relations inter-produits, mais plutôt comme dans le chapitre 6 avec une approche "chaque produit contre le reste des autres", plutôt que "chaque couple de produit l'un contre l'autre".

1. Il existe des méthodes pour obtenir des graphes de produits similaires, utilisés par exemple pour de la recommandation

2. Quoique le chapitre 5 semble montrer que munir les catégories d'une relation d'ordre est pertinent.

Mesure de la cannibalisation Une des limites majeures du chapitre 6 est que l'on suppose qu'il y a une cannibalisation totale des ventes . Une augmentation des ventes se fait forcément au détriment des ventes d'un autre produit. Or, ce n'est pas tout le temps le cas, puisqu'une augmentation de la compétitivité peut aussi se traduire par une augmentation des ventes. On pourrait proposer le modèle suivant, qui étend le modèle de ce chapitre :

$$\begin{cases} x_{i,t} &= \epsilon_{i,t}(\lambda_{i,t}) \\ \lambda_{i,t} &= s(t) \dfrac{\phi(x_{i,t-1}/s(t-1), \theta_{i,t})}{\left(1 + \sum\limits_{j \in [1,d]} \phi(x_{j,t-1}/s(t-1), \theta_{j,t})\right)^{\gamma}} \end{cases}$$

On introduit un coefficient γ compris entre 0 et 1 qui est en quelque sorte une mesure de la cannibalisation. Lorsque $\gamma = 1$, on est dans un cas de cannibalisation totale, c'est à dire comme dans le chapitre 6. Au contraire, lorsque $\gamma = 0$, il n'y a pas de cannibalisation, et les ventes sont indépendantes. On se retrouve alors dans une approche auto-régressive telle que définie aux chapitres 4 et 5.

Valider le choix d'un tel γ est une question ouverte. Une approche par hold-out est bien sur possible, même si ce type d'approche peut être assez coûteux en terme de temps de calcul. Des approches par minimisation du risque empirique peuvent aussi être tentées. Le calcul d'un tel coefficient pourrait être très intéressant du point de vue de l'explicabilité, car on pourrait ainsi avoir une mesure la tendance de telles ou telle catégorie de produits à se cannibaliser, ce qui peut avoir de nombreux intérêts pour piloter une marketplace. Ainsi, on pourrait évaluer mieux estimer l'impact de la promotion d'un produit sur les ventes de toute sa catégorie. Comme expliqué au chapitre 4, on peut aussi déterminer les catégories pour lesquelles les prédictions sont les plus intéressantes, puisque l'effet d'une rupture de stocks sur des catégories avec forte cannibalisation sera moins important.

Modèle de séries concurrentes à base d'arbres Le choix de modèle neuronaux au chapitre 6 est surtout lié à la facilité avec laquelle on peut différencier des réseaux de neurones . Cependant, les approches régressives les plus efficaces en pratique, sont des approches à base d'arbres et de forêts similaires à celle du chapitre 5. Elles sont également plus pratique, en permettant d'intégrer des données catégorielles et discrètes plus aisément. Il semble possible d'adapter les idées du chapitre 6 à des modèles d'arbres, en proposant des modèles d'arbres (ou de forêts) concurrentes.

Une difficulté majeure pour l'application directe de ce type d'idée aux arbres est le fait que la modification du poids produit par une feuille modifie l'ensemble des prédictions, et non pas seulement les prédictions associés à cette feuille. Cependant, pour des arbres CART, et si on abandonne la modélisation par loi de Poisson pour la remplacer par des pertes L1, cet effet n'est pas gênant pour déterminer comment séparer les prédictions associés à une feuille, puisque toutes les prédictions affectées aux autres feuilles sont impactés par un même facteur multiplicatif. Cela permet de simplifier le problème, car tout se passe alors comme si on ajoutait une pénalité chaque fois que l'on sépare un noeud, cette pénalité variable pouvant être calculé sans faire appel à ce qu'il y a en dehors du noeuds. L'ajout de cet étape est complètement non standard et ne permet pas d'utiliser les algorithmes d'arbre existants. Il serait donc nécessaire d'adapter l'algorithme pour CART, ce qui donnerait ensuite des variantes pour les algorithmes de Random Forest ou de Boosting.

Bibliographie

[AC10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4 :40–79, 2010. 17

[ACKR16] Arianna Agosto, Giuseppe Cavaliere, Dennis Kristensen, and Anders Rahbek. Modeling corporate defaults : Poisson autoregressions with exogenous covariates (parx). *Journal of Empirical Finance*, 38 :640–663, 2016. 14, 74, 75

[ADF19] Pierre Alquier, Paul Doukhan, and Xiequan Fan. Exponential inequalities for nonstationary markov chains. *Dependence Modeling*, 7(1) :150–168, 2019. 16, 29, 75, 78

[Ahl21] Thomas D Ahle. Sharp and simple bounds for the raw moments of the binomial and poisson distributions. *arXiv preprint arXiv :2103.17027*, 2021. XVII

[ALW13] Pierre Alquier, Xiaoyin Li, and Olivier Wintenberger. Prediction of time series by statistical learning : general losses and fast rates. *Dependence Modeling*, 1(2013) :65–93, 2013. 15

[AQS01] Ilan Alon, Min Qi, and Robert J Sadowski. Forecasting aggregate retail sales : : a comparison of artificial neural networks and traditional methods. *Journal of retailing and consumer services*, 8(3) :147–156, 2001. 71

[AW12a] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3) :883–913, 2012. 17

[AW+12b] Pierre Alquier, Olivier Wintenberger, et al. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3) :883–913, 2012. 15

[BBC+01] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8) :1200–1211, 2001. 30

[BBL00] Peter L. Bartlett, Stephane Boucheron, and Gábor Lugosi. Model Selection and Error Estimation. *SSRN Electronic Journal*, 2000. 35

[BBL02] Peter Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48 :85–113, 2002. 4, 19, 24

[BBL03] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003. 12

[BBO17] Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. Conditional time series forecasting with convolutional neural networks. *stat*, 1050 :16, 2017. 62, 66

[BC19] Patrice Bertail and Gabriela Ciolek. New Bernstein and Hoeffding type inequalities for regenerative Markov chains. *Latin American Journal of Probability and Mathematical Statistics*, 16(1) :259, 2019. 29

[BCC21] Júlio Barros, Paulo Cortez, and M Sameiro Carvalho. A systematic literature review about dimensioning safety stock under uncertainties and risks in the procurement process. *Operations Research Perspectives*, page 100192, 2021. 54

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4) :929–965, 1989. 12

[BHZ09] Zou Bin, Zhang Hai, and Xu Zongben. Learning from uniformly ergodic markov chains. *Journal of complexity*, 25 :188–200, 2009. 18

[BJRL15] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis : forecasting and control*. John Wiley & Sons, 2015. 72

[BLP95] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica : Journal of the Econometric Society*, pages 841–890, 1995. 72, 87

[BLP04] Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data : The new car market. *Journal of political Economy*, 112(1) :68–105, 2004. 72

[BM06] G. Blanchard and Pascal Massart. Discussion : Local rademacher complexities and oracle inequalities in risk minimization. *Annals of statistics*, 34 :2664–2671, 2006. 17, 26

[BM12] Anna-Lena Beutel and Stefan Minner. Safety stock planning under causal demand forecasting. *International Journal of Production Economics*, 140(2) :637–645, 2012. 54

[BMW01] Suman Basuroy, Murali K Mantrala, and Rockney G Walters. The impact of category management on retailer prices and performance : Theory and evidence. *Journal of Marketing*, 65(4) :16–32, 2001. 72

[BN92] Prabir Burman and Deborah Nolan. Data dependent estimation of prediction functions. *Journal of time series analysis*, 13 (3) :189–207, 1992. 17

[BOL05] Stéphane Boucheron, Bousquet Olivier, and Gábor Lugosi. Theory of classification : A survey of some recent advances. *ESAIM : PS*, 9 :323–375, 2005. 26

[BSB+19] Kasun Bandara, Peibei Shi, Christoph Bergmeir, Hansika Hewamalage, Quoc Tran, and Brian Seaman. Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *Neural Information Processing*, pages 462–474. Springer International Publishing, 2019. 61, 62, 69, 71, 73

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. 15

[CC+04] Bo-Juen Chen, Ming-Wei Chang, et al. Load forecasting using support vector machines : A study on eunite competition 2001. *IEEE transactions on power systems*, 19(4) :1821–1830, 2004. 62

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016. 67

[Cha14] Nicolas Chapados. Effective bayesian modeling of groups of related count time series. In *International conference on machine learning*, pages 1395–1403. PMLR, 2014. 57, 61, 62

[CM+91] C-K Chu, James Stephen Marron, et al. Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics*, 19(4) :1906–1918, 1991. 18

[Com15] Richard Combes. An extension of McDiarmid's inequality. *arXiv :1511.05240 [cs, math, stat]*, 2015. arXiv : 1511.05240. 30, 44, XIV

[CPC19] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability : A survey on methods and metrics. *Electronics*, 8(8) :832, 2019. 55

[CTM20] Vitor Cerqueira, Luis Torgo, and Igor Mozetic. Evaluating time series forecasting models : an empirical study on performance estimation methods. *Machine Learning*, 109 :1997–2028, 2020. 17

[CW16] Likai Chen and Wei Biao Wu. Stability and asymptotics for autoregressive processes. *Electronic Journal of Statistics*, 10(2) :3723–3751, 2016. 5

[CW18] Likai Chen and Wei Biao Wu. Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research*, 18(231) :1–46, 2018. 16, 29, 40

[DDF19] Jérôme Dedecker, Paul Doukhan, and Xiequan Fan. Deviation inequalities for separately Lipschitz functionals of composition of random functions. *Journal of Mathematical Analysis and Applications*, 479(2) :1549–1568, 2019. 14, 16, 29

[DDL+07] Jérôme Dedecker, Paul Doukhan, Gabriel Lang, León R José Rafael, Sana Louhichi, and Clémentine Prieur. Weak dependence. In *Weak dependence : With examples and applications*, pages 9–20. Springer, 2007. 15

[DF15] Jerome Dedecker and Xieqan Fan. Deviation inequalities for separately lipschitz functionals of iterated random functions. *Stochastic Processes and their Applications*, 125 :60–90, 2015. 14, 16, 29, 32, 33, 35, 75

[DFT12] Paul Doukhan, Konstantinos Fokianos, and Dag Tjøstheim. On weak dependence conditions for poisson autoregressions. *Statistics & Probability Letters*, 82(5) :942–948, 2012. 75

[DJ94] David L. Donoho and Jain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994. 14

[DNT20] Paul Doukhan, Michael H. Neumann, and Lionel Truquet. Stationarity and ergodic properties for some observation-driven models in random environments, 2020. 74

[DPRS17] Paul Doukhan, Denys Pommeret, Joseph Rynkiewicz, and Yahia Salhi. A class of random field memory models for mortality forecasting. *Insurance : Mathematics and Economics*, 77 :97–110, 2017. 30

[DT07] Paul Doukhan and Lionel Truquet. A fixed point approach to model random fields. *Alea*, 3 :111–132, 2007. 5, 16, 30, 31, 32, 40, 41

[DT20] Max Zinsou Debaly and Lionel Truquet. Iterations of dependent random maps and exogeneity in nonlinear dynamics, 2020. 14, 75

[DZZ16] Richard A Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4) :1077–1096, 2016. 71

[EA07] Volkan Ş Ediger and Sertac Akar. Arima forecasting of primary energy demand by fuel in turkey. *Energy policy*, 35(3) :1701–1708, 2007. 62

[Fok12] Konstantinos Fokianos. Count time series models. In *Handbook of statistics*, volume 30, pages 315–347. Elsevier, 2012. 74

[FSTD20] Konstantinos Fokianos, Bård Støve, Dag Tjøstheim, and Paul Doukhan. Multivariate count autoregression. *Bernoulli*, 26(1) :471–499, Feb 2020. 6, 74

[FT17] Konstantinos Fokianos and Lionel Truquet. On categorical time series models with covariates. *Stochastic Processes and their Applications*, 129, 09 2017. 59, 74

[Gar21] Rémy Garnier. Concurrent neural network : a model of competition between times series. *Annals of Operations Research*, pages 1–20, 2021. 6

[GB19] Rémy Garnier and Arnaud Belletoile. Une approche multi-séries pour la prévision de la demande sur des données d'e-commerce. *APIA*, page 40, 2019. 6, 71

[Git] Concurrent neural network. https://github.com/garnier94/Concurrent_Neural_Network. 79

[Has17] Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017. 66

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301) :13–30, 1963. 12

[HS14] Hanyuan Hang and Ingo Steinwart. Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127 :184–199, 2014. 15

[Kon14] Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 28–36, Bejing, China, 22–24 Jun 2014. PMLR. 44

[KP10] Mahesh Kumar and Nitin R Patel. Using clustering to improve sales forecasts in retail merchandising. *Annals of Operations Research*, 174(1) :33–46, 2010. 72

[KPW02] Mahesh Kumar, Nitin R Patel, and Jonathan Woo. Clustering seasonality patterns in the presence of errors. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 557–563. ACM, 2002. 63

[KR08] Leonid (Aryeh) Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6) :2126–2158, 2008. 29

[KT19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 29

[Kut02] Samuel Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. *Department Computer Science, University of Chicago, Chicago, IL. Technical report TR-2002-04*, 2002. 44

[Lin89] Winston T Lin. Modeling and forecasting hospital patient movements : Univariate and multiple time series approaches. *International Journal of Forecasting*, 5(2) :195–208, 1989. 71

[Lin02] Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002. 35

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. 58

[Lug02a] G. Lugosi. *Pattern Classification and Learning Theory*, pages 1–56. Springer Vienna, Vienna, 2002. 16, 35, 40

[Lug02b] Gábor Lugosi. *In Györfi, L., editor, Principles of nonparametric learning*, chapter Pattern classification and learning theory. Springer-Verlag, 2002. 20

[Lüt05] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005. 71

[M+89] Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1) :148–188, 1989. 44, 46

[Mas03] Pascal Massart. *Ecole d'été de probabilité de Saint-Flour XXXIII*. Springer-Verlag, 2003. 26

[MH19] John Miller and Moritz Hardt. Stable recurrent models. *In Proceedings of ICLR 2019*, 2019. 76

[MK17] M. Mohri and V. Kuznetsov. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106 :93–117, 2017. 17

[MR10] Mehryar Mohri and Afshin Rostamizedeh. Stability bounds for stationnary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11 :789–814, 2010. 15, 17

[Pau15] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79) :32 p., 2015. 15, 20, 21, 29, IV

[PG11] Amandine Pierrot and Yannig Goude. Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, 2011, 2011. 62

[PR19] Rasmus Søndergaard Pedersen and Anders Rahbek. Testing garch-x type models. *Econometric Theory*, 35(5) :1012–1047, 2019. 74

[RCS20] Shuyun Ren, Hau-Ling Chan, and Tana Siqin. Demand forecasting in retail operations for fashionable products : methods, practices, and real case study. *Annals of Operations Research*, 291(1) :761–777, 2020. 71

[Rel] Relex. https://www.relexsolutions.com/fr/. 68

[Ros56] Murray Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1) :43, 1956. 15

[RR04] Gareth O. Roberts and Jeffrey .S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Survey*, 1 :20–71, 2004. 15

[SAH+21] Evangelos Spiliotis, Mahdi Abolghasemi, Rob J Hyndman, Fotios Petropoulos, and Vassilios Assimakopoulos. Hierarchical forecast reconciliation with machine learning. *Applied Soft Computing*, 112 :107756, 2021. 57

[SP97] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11) :2673–2681, 1997. 29

[SRG05] Sundara Raghavan Srinivasan, Sreeram Ramakrishnan, and Scott E Grasman. Identifying the effects of cannibalization on the product portfolio. *Marketing intelligence & planning*, 2005. 71, 72

[SSF16] Matthias W Seeger, David Salinas, and Valentin Flunkert. Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems*, pages 4646–4654, 2016. 62

[Tas00] Leonard Tashman. Out-of-sample tests of forecasting accuracy : an analysis and review. *International journal of forecasting*, 16 :437–450, 2000. 17

[Tay03] James W Taylor. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8) :799–805, 2003. 62

[TC01] Francis EH Tay and Lijuan Cao. Application of support vector machines in financial time series forecasting. *omega*, 29(4) :309–317, 2001. 71

[TKF15] Juan R Trapero, Nikolaos Kourentzes, and Robert Fildes. On the identification of sales forecasting models in the presence of promotions. *Journal of the operational Research Society*, 66(2) :299–307, 2015. 61

[Van21] Nicolas Vandeput. *Data science for supply chain forecasting*. De Gruyter, 2021. 53, 71, 72, 79

[Vap99] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5) :988–999, 1999. 11

[VHSD10] Harald J Van Heerde, Shuba Srinivasan, and Marnik G Dekimpe. Estimating cannibalization rates for pioneering innovations. *Marketing Science*, 29(6) :1024–1039, 2010. 72

[VS18] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks : analysis and efficient estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3835–3844. Curran Associates, Inc., 2018. 33

[War16] Lutz Warnke. On the method of typical bounded differences. *Combinatorics, Probability and Computing*, 25(2) :269–299, 2016. 44

[WK19] Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. *COLT 2019 : Proceedings of Machine Learning Research*, 99 :1–40, 2019. 15, 24, 26

[Yel10] Phillip M Yelland. Bayesian forecasting of parts demand. *International Journal of Forecasting*, 26(2) :374–396, 2010. 61, 62

[ZBH+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communication of the ACM*, 64(3) :107–115, 2021. 18

[ZWC18] Bing Zhang, Jhen-Long Wu, and Pei-Chann Chang. A multiple time series-based recurrent neural network for short-term load forecasting. *Soft Computing*, 22(12) :4099–4112, 2018. 71

Liste des figures

Liste des tableaux

Annexe A

Appendices du chapitre 2

A.1 Preuve du théorème 2.1

We prove (2.14) with the sign +, the proof for the sign − is symmetric. Since $0 < \varepsilon \leq 1$, if $b = \lfloor \frac{m\varepsilon^2}{1+9\ln(2)} \rfloor$, then $\varepsilon - \frac{b}{m} > 0$. Using the proposition 2.5, we have :

$$\mathbb{P}\left(\pm(\hat{\mathrm{L}}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n)) > \varepsilon\right) \leq \exp\left(-2\frac{(m-b)(\varepsilon - \frac{b}{m})^2}{9t_{mix}}\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right).$$

Now, if $\tilde{b} = \frac{m\varepsilon^2}{1+9\ln(2)}$, then $\tilde{b} - 1 \leq b \leq \tilde{b}$, and we get

$$\mathbb{P}\left(\pm(\hat{\mathrm{L}}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n)) > \varepsilon\right) \leq \exp\left(-2\frac{(m-\tilde{b})(\varepsilon - \frac{\tilde{b}}{m})^2}{9t_{mix}}\right) + 2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\exp\left(-\frac{\tilde{b}\ln(2)}{t_{mix}}\right).$$

Moreover,

$$\begin{aligned}
&\exp\left(-2\frac{(m-\tilde{b})(\varepsilon - \frac{\tilde{b}}{m})^2}{9t_{mix}}\right) + 2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\exp\left(-\frac{\tilde{b}\ln(2)}{t_{mix}}\right) = \\
&\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + \exp\left(-m\left(\frac{2(1-\frac{\tilde{b}}{m})(\varepsilon - \frac{\tilde{b}}{m})^2}{9t_{mix}} - \frac{\tilde{b}\ln(2)}{mt_{mix}}\right)\right)\right)\exp\left(-\frac{\tilde{b}\ln(2)}{t_{mix}}\right) = \\
&\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + \exp\left(-\frac{m}{t_{mix}}\left(\frac{2(1-\frac{\tilde{b}}{m})(\varepsilon - \frac{\tilde{b}}{m})^2}{9} - \frac{\tilde{b}\ln(2)}{m}\right)\right)\right)\exp\left(-\frac{\tilde{b}\ln(2)}{t_{mix}}\right).
\end{aligned}$$

Now,

$$\begin{aligned}
&\frac{2(1-\frac{\tilde{b}}{m})(\varepsilon - \frac{\tilde{b}}{m})^2}{9} - \frac{\tilde{b}\ln(2)}{m} = \frac{2(1-\frac{\varepsilon^2}{1+9\ln(2)})(\varepsilon - \frac{\varepsilon^2}{1+9\ln(2)})^2}{9} - \frac{\varepsilon^2\ln(2)}{1+9\ln(2)} \geq \\
&\varepsilon^2\left(\frac{2(1-\frac{1}{1+9\ln(2)})(1-\frac{1}{1+9\ln(2)})^2}{9} - \frac{\ln(2)}{1+9\ln(2)}\right) = \varepsilon^2\left(\frac{2}{9}\left(1 - \frac{1}{1+9\ln(2)}\right)^3 - \frac{\ln(2)}{1+9\ln(2)}\right) > 0,
\end{aligned}$$

and finally

$$\begin{aligned}
&\exp\left(-2\frac{(m-\tilde{b})(\varepsilon - \frac{\tilde{b}}{m})^2}{9t_{mix}}\right) + 2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\exp\left(-\frac{\tilde{b}\ln(2)}{t_{mix}}\right) \leq \\
&\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + 1\right)\exp\left(-\frac{\tilde{b}\ln(2)}{t_{mix}}\right) = \left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + 1\right)\exp\left(-\frac{m\varepsilon^2\ln(2)}{(1+9\ln(2))t_{mix}}\right).
\end{aligned}$$

■

A.2 Preuve de la Proposition 2.6

First, we will prove that for a stationary uniformly ergodic Markov chain $(X_t)_{t\in\mathbb{Z}}$ and any function $g \in \mathscr{G}$:

$$P\left(\frac{1}{1+a}\frac{1}{m}\sum_{k=1}^{m} L(g(X_k)) - \mathbb{E}(L(g(X_0))) > \varepsilon\right) \leq \exp\left(-mD_a^+\varepsilon\right), \tag{A.1}$$

where $\mathbb{E}(L(g(X_0)))$ is computed under the stationary law of $(X_t)_{t\in\mathbb{Z}}$.

Since $0 \leq L(g(X_0)) \leq 1$, $0 \leq L(g(X_0))^2 \leq L(g(X_0)) \leq 1$, and $V(L(g(X_0))) \leq \mathbb{E}(L(g(X_0)))(1-\mathbb{E}(L(g(X_0)))) \leq \mathbb{E}(L(g(X_0)))$. Moreover,

$$P\left(\frac{1}{1+a}\frac{1}{m}\sum_{k=1}^{m} L(g(X_k)) - \mathbb{E}(L(g(X_0))) > \varepsilon\right) =$$
$$P\left(\frac{1}{m}\sum_{k=1}^{m} L(g(X_k)) - \mathbb{E}(L(g(X_0))) > a\mathbb{E}(L(g(X_0))) + (1+a)\varepsilon\right).$$

Let $t = a\mathbb{E}(L(g(X_0))) + (1+a)\varepsilon$, by the theorem 3.4 of [Pau15], we have :

$$P\left(\frac{1}{m}\sum_{k=1}^{m} L(g(X_k)) > t\right) \leq \exp\left(-\frac{m^2t^2\gamma_{ps}}{8(m+1/\gamma_{ps})\mathbb{E}(L(g(X_0))) + 20mt}\right) =$$
$$\exp\left(-\frac{mt^2\gamma_{ps}}{8(1+\frac{1}{m\gamma_{ps}})\mathbb{E}(L(g(X_0))) + 20t}\right)$$

Now,

$$8(1+\frac{1}{m\gamma_{ps}})\mathbb{E}(L(g(X_0))) + 20t \leq t\left(\frac{8(1+1/\gamma_{ps})}{a} + 20\right),$$

hence

$$P\left(\frac{1}{m}\sum_{k=1}^{m} L(g(X_k)) > t\right) \leq \exp\left(-\frac{mt\gamma_{ps}}{\frac{8(1+1/\gamma_{ps})}{a} + 20}\right) \leq$$
$$\exp\left(-\frac{mat\gamma_{ps}}{8(1+1/\gamma_{ps}) + 20}\right) \leq \exp\left(-\frac{ma(1+a)\varepsilon\gamma_{ps}}{8(1+1/\gamma_{ps}) + 20}\right),$$

and we deduce equation (A.1).

Now, using equation (3.27) of Paulin [Pau15], we get

$$P\left(\frac{1}{1+a}\frac{1}{m-b}\sum_{k=n+b+1}^{n+m} L(\hat{g}_1^n(X_k)) - \mathbb{L}(\hat{g}_1^n) > \varepsilon\right) \leq \exp\left(-(m-b)\gamma_{ps}D_a^+\varepsilon\right) + 2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right).$$

Finally, by the same arguments as in the proof of proposition 2.5, we get equation (2.23). The proof of equation (2.24) is symmetric. ■

A.3 Preuve du Théorème 2.5

We prove (2.25), the proof for (2.26) is symmetric. Let us define $\tilde{b} = \frac{D_a^+\varepsilon m}{4\ln(2)}$. Since $\lfloor\tilde{b}\rfloor > \tilde{b} - 1$, equation (2.23) yields

$$P\left(\frac{1}{1+a}\hat{L}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) > \varepsilon\right) \leq$$
$$\exp\left(-\gamma_{ps}D_a^+\left(m - \frac{\tilde{b}}{m}\right)\left(\varepsilon - \frac{\tilde{b}}{m}\right)\right) + 2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\exp\left(-\frac{\tilde{b}\ln(2)}{t_{mix}}\right) =$$
$$\exp\left(-\gamma_{ps}D_a^+m\varepsilon\left(1 - \frac{D_a^+\varepsilon}{4\ln(2)}\right)\left(1 - \frac{D_a^+}{4\ln(2)}\right)\right) + 2\exp\left(\frac{\ln(2)}{t_{mix}}\right)\exp\left(-\frac{D_a^+\varepsilon m}{4t_{mix}}\right) =$$
$$\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + \exp\left(-\gamma_{ps}D_a^+m\varepsilon\left(\left(1 - \frac{D_a^+\varepsilon}{4\ln(2)}\right)\left(1 - \frac{D_a^+}{4\ln(2)}\right) - \frac{1}{4t_{mix}\gamma_{ps}}\right)\right)\right)\exp\left(-\frac{D_a^+\varepsilon m}{4t_{mix}}\right).$$

Now, since $\varepsilon \leq 1$, $\gamma_{ps} \geq \frac{1}{2t_{mix}} \Leftrightarrow 2 \geq \frac{1}{2t_{mix}\gamma_{ps}}$, we get

$$\mathrm{P}\left(\frac{1}{1+a}\hat{\mathrm{L}}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) > \varepsilon\right) \leq$$
$$\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + \exp\left(-\gamma_{ps}\mathrm{D}_a^+ m\varepsilon\left(\left(1 - \frac{\mathrm{D}_a^+}{4\ln(2)}\right)^2 - \frac{1}{2}\right)\right)\right)\exp\left(-\frac{\mathrm{D}_a^+\varepsilon m}{4t_{mix}}\right).$$

Finally, noting that $\mathrm{D}_a^+ = \dfrac{a(1+a)}{8\left(1+\frac{1}{\gamma_{ps}}\right)+20} \leq \dfrac{1}{14}$, we have

$$\left(1 - \frac{\mathrm{D}_a^+}{4\ln(2)}\right)^2 - \frac{1}{2} \geq 0,$$

and

$$\mathrm{P}\left(\frac{1}{1+a}\hat{\mathrm{L}}_m(\hat{g}_1^n) - \mathbb{L}(\hat{g}_1^n) > \varepsilon\right) \leq \left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + 1\right)\exp\left(-\frac{\mathrm{D}_a^+\varepsilon m}{4t_{mix}}\right)$$

■

A.4 Preuve de la Proposition 2.7

Let us considere a finite collection of functions $\{g_1, \cdots, g_{\mathrm{N}}\}$. For any integers b and m, with $0 \leq b < m$, and a sample $(\mathrm{X}_1, \cdots, \mathrm{X}_{m+b})$ where the m last variables $(\mathrm{X}_{b+1}, \cdots, \mathrm{X}_{m+b})$ follow the stationary law of the Markov chain $(\mathrm{X}_t)_{t\in\mathbb{Z}}$. Let us define :

$$g_{\hat{k}} = \min_{k\in\{1,\cdots,\mathrm{N}\}} \frac{1}{m+b}\sum_{t=1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) \text{ and } g_{\tilde{k}} = \min_{k\in\{1,\cdots,\mathrm{N}\}} \mathbb{L}(g_k).$$

We will give bounds involving the m last variables : $\hat{\mathrm{L}}_m(g_k) := \dfrac{1}{m}\sum_{t=b+1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t))$. Note that, for any function g_k :

$$\frac{1}{m+b}\sum_{t=1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) - \frac{1}{m}\sum_{t=b+1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) \leq$$
$$\frac{1}{m+b}\sum_{t=1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) - \frac{1}{m+b}\sum_{t=b+1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) \leq \frac{b}{m+b}, \tag{A.2}$$

and

$$\frac{1}{m}\sum_{t=b+1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) - \frac{1}{m+b}\sum_{t=1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t) \leq$$
$$\frac{m+b}{m(m+b)}\sum_{t=b+1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) - \frac{m}{m(m+b)}\sum_{t=1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) \leq$$
$$\frac{b}{m(m+b)}\sum_{t=b+1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) \leq \frac{b}{m+b}, \tag{A.3}$$

so

$$\frac{1}{m}\sum_{t=b+1}^{m+b} \mathrm{L}(g_{\hat{k}}(\mathrm{X}_t)) - \frac{1}{m}\sum_{t=b+1}^{m+b} \mathrm{L}(g_k(\mathrm{X}_t)) \leq \frac{2b}{m+b}. \tag{A.4}$$

By the lemma 2.2 and the union bound, with probability at least $1-\delta$, for all $k \in \{1, \cdots, \mathrm{N}\}$,

$$\mathbb{L}\left(g_k\right) - \mathbb{L}(g^*) \leq$$
$$\hat{\mathrm{L}}_m\left(g_k\right) - \hat{\mathrm{L}}_m(g^*) + \sqrt{\frac{8(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps}m}} \times \omega\left(\mathbb{L}(g_k) - \mathbb{L}(g^*)\right) + \frac{40\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps}m},$$

and

$$\mathbb{L}(g^*) - \mathbb{L}\left(g_{\tilde{k}}\right) \leq$$
$$\hat{\mathrm{L}}_m(g^*) - \hat{\mathrm{L}}_m\left(g_{\tilde{k}}\right) + \sqrt{\frac{8(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}} \times \omega\left(\mathbb{L}\left(g_{\tilde{k}}\right) - \mathbb{L}(g^*)\right) + \frac{40\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}.$$

Since $\mathbb{L}\left(g_{\tilde{k}}\right) - \mathbb{L}(g^*) \leq \mathbb{L}\left(g_k\right) - \mathbb{L}(g^*)$, for any $k \in \{1, \cdots, \mathrm{N}\}$, by summing the two inequalities, we obtain

$$\mathbb{L}\left(g_k\right) - \mathbb{L}\left(g_{\tilde{k}}\right) \leq \hat{\mathrm{L}}_m\left(g_k\right) - \hat{\mathrm{L}}_m\left(g_{\tilde{k}}\right) +$$
$$2\sqrt{\frac{8(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}} \times \omega\left(\mathbb{L}\left(g_{\tilde{k}}\right) - \mathbb{L}(g^*)\right) + \frac{80\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}.$$

As $\hat{\mathrm{L}}_m\left(g_{\hat{k}}\right) - \hat{\mathrm{L}}_m\left(g_{\tilde{k}}\right) \leq \frac{2b}{m+b}$, with probability larger than $1-\delta$,

$$\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}\left(g_{\tilde{k}}\right) \leq$$
$$\frac{2b}{m+b} + 2\sqrt{\frac{8(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}} \times \omega\left(\mathbb{L}\left(g_{\tilde{k}}\right) - \mathbb{L}(g^*)\right) + \frac{80\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}.$$

Let τ^*_m be defined as the statement of the theorem. If $\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*) \geq \tau^*_{m+b}$, then $\frac{\omega\left(\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)\right)}{\sqrt{m+b}} \leq \sqrt{\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)\tau^*_{m+b}}$, and we have

$$\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}\left(g_{\tilde{k}}\right) \leq$$
$$\frac{2b}{m+b} + 2\sqrt{\frac{8(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps}}} \times \sqrt{\frac{m+b}{m}}\sqrt{\tau^*_{m+b}}\sqrt{\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)} + \frac{80\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}. \tag{A.5}$$

For $0 < \theta < 1$ we have :

$$\frac{\theta^2}{2}\left(\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)\right) - 2\sqrt{\frac{8(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps}}} \times \sqrt{\frac{m+b}{m}}\sqrt{\tau^*_{m+b}}\sqrt{\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)}\theta +$$
$$\frac{16(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps}} \frac{m+b}{m}\tau^*_{m+b} =$$
$$\left(\frac{\theta}{\sqrt{2}}\sqrt{\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)} - \sqrt{\frac{16(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps}}} \times \sqrt{\frac{m+b}{m}}\sqrt{\tau^*_{m+b}}\right)^2 \geq 0,$$

and

$$2\sqrt{\frac{8(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps}}} \times \sqrt{\frac{m+b}{m}}\sqrt{\tau^*_{m+b}}\sqrt{\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)} \leq$$
$$\frac{\theta}{2}\left(\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)\right) + \frac{16(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\theta\gamma_{ps}} \frac{m+b}{m}\tau^*_{m+b},$$

so

$$\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}\left(g_{\tilde{k}}\right) \leq \frac{\theta}{2}\left(\mathbb{L}\left(g_{\hat{k}}\right) - \mathbb{L}(g^*)\right) + \frac{2b}{m+b} + \frac{16(1+\frac{1}{\gamma_{ps}})\log\left(\frac{\mathrm{N}}{\delta}\right)}{\theta\gamma_{ps}} \frac{m+b}{m}\tau^*_{m+b} + \frac{80\log\left(\frac{\mathrm{N}}{\delta}\right)}{\gamma_{ps} m}.$$

Hence, with probability larger than $1-\delta$

$$\left(1-\frac{\theta}{2}\right)\left(\mathbb{L}\left(g_{\hat{k}}\right)-\mathbb{L}(g^*)\right) \leq \mathbb{L}\left(g_{\tilde{k}}\right)-\mathbb{L}(g^*) + \frac{2b}{m+b} + \frac{16(1+\frac{1}{\gamma_{ps}})\log\left(\frac{N}{\delta}\right)}{\theta\gamma_{ps}}\frac{m+b}{m}\tau^*_{m+b} + \frac{80\log\left(\frac{N}{\delta}\right)}{\gamma_{ps}m},$$

and

$$\left(\mathbb{L}\left(g_{\hat{k}}\right)-\mathbb{L}(g^*)\right) \leq \frac{1}{1-\frac{\theta}{2}}\left(\mathbb{L}\left(g_{\tilde{k}}\right)-\mathbb{L}(g^*) + \frac{2b}{m+b} + \frac{16(1+\frac{1}{\gamma_{ps}})\log\left(\frac{N}{\delta}\right)}{\theta\gamma_{ps}}\frac{m+b}{m}\tau^*_{m+b} + \frac{80\log\left(\frac{N}{\delta}\right)}{\gamma_{ps}m}\right). \tag{A.6}$$

Since, for $0<\theta<1$, $\frac{1}{1-\frac{\theta}{2}} \leq 1+\theta$, we get, with probability larger than $1-\delta$:

$$\mathbb{L}\left(g_{\hat{k}}\right)-\mathbb{L}(g^*) \leq (1+\theta)\times\left(\mathbb{L}\left(g_{\tilde{k}}\right)-\mathbb{L}(g^*) + \frac{2b}{m+b} + \frac{16(1+\frac{1}{\gamma_{ps}})\log\left(\frac{N}{\delta}\right)}{\theta\gamma_{ps}}\frac{m+b}{m}\tau^*_{m+b} + \frac{80\log\left(\frac{N}{\delta}\right)}{\gamma_{ps}m}\right),$$

or

$$\mathbb{L}\left(g_{\hat{k}}\right)-\mathbb{L}(g^*)-(1+\theta)\left(\mathbb{L}\left(g_{\tilde{k}}\right)-\mathbb{L}(g^*)+\frac{2b}{m+b}\right) \leq (1+\theta)\left(\frac{16(1+\frac{1}{\gamma_{ps}})(m+b)\tau^*_{m+b}+80\theta}{\theta\gamma_{ps}m}\log\left(\frac{N}{\delta}\right)\right). \tag{A.7}$$

Note that we have done the reasoning if $\mathbb{L}\left(g_{\hat{k}}\right)-\mathbb{L}(g^*) \geq \tau^*_{m+b}$. However, if $\mathbb{L}\left(g_{\hat{k}}\right)-\mathbb{L}(g^*) < \tau^*_{m+b}$, the bound (A.7) is obvious. Now, we deduce from equation (A.7) that

$$\mathbb{P}\left(\mathbb{L}\left(g_{\hat{k}}\right)-\mathbb{L}(g^*)-(1+\theta)\left(\mathbb{L}\left(g_{\tilde{k}}\right)-\mathbb{L}(g^*)+\frac{2b}{m+b}\right)>\varepsilon\right) \leq N\exp\left(-\frac{1}{1+\theta}\frac{\theta\gamma_{ps}m}{16(1+\frac{1}{\gamma_{ps}})(m+b)\tau^*_{m+b}+80\theta}\varepsilon\right).$$

Now, considering the actual chain $(X_t)_{t\in\mathbb{N}}$, in the framework of section 2.2.4, we get for any realization $x_1,\cdots,x_n$ of $X_1,\cdots,X_n$, integers b and m, with $0 \leq b < m$:

$$\mathbb{P}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)-(1+\theta)\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)>\varepsilon+\frac{(1+\theta)2b}{m}\right) \leq N\exp\left(-\frac{1}{1+\theta}\frac{\theta\gamma_{ps}(m-b)}{16(1+\frac{1}{\gamma_{ps}})m\tau^*_m+80\theta}\varepsilon\right)+2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right). \tag{A.8}$$

■

A.5 Preuve du Théorème 2.8

Applying the proposition 2.7 to $2(1+\theta)\varepsilon$, we get

$$\mathbb{P}\left(\frac{\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right)-\mathbb{L}(g^*)}{2(1+\theta)}-\frac{\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right)-\mathbb{L}(g^*)\right)}{2}>\varepsilon+\frac{b}{m}\right) \leq N\exp\left(-2\frac{\theta\gamma_{ps}(m-b)}{16(1+\frac{1}{\gamma_{ps}})m\tau^*_m+80\theta}\varepsilon\right)+2\exp\left(-\frac{b\ln(2)}{t_{mix}}\right). \tag{A.9}$$

Let us define $\tilde{b} = \dfrac{\theta m \varepsilon}{(16(1+\frac{1}{\gamma_{ps}})m\tau^*_m + 80\theta)2\ln(2)}$. Following the same reasoning as in the proof of theorem 2.5, we get

$$\mathbb{P}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right) - \mathbb{L}(g^*) - (1+\theta)\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right) - \mathbb{L}(g^*)\right) > \varepsilon\right) \leq$$
$$\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + N\exp\left(-\frac{2\theta\gamma_{ps} m\varepsilon}{(16(1+\frac{1}{\gamma_{ps}})m\tau^*_m + 80\theta)}\left(\left(1 - \frac{\theta}{(16(1+\frac{1}{\gamma_{ps}})m\tau^*_m + 80\theta)2\ln(2)}\right)^2 - \frac{1}{2}\right)\right)\right) \times$$
$$\exp\left(-\frac{\theta\gamma_{ps}\varepsilon m}{(16(1+\frac{1}{\gamma_{ps}})m\tau^*_m + 80\theta)2t_{mix}}\right).$$

Noting that $\dfrac{\theta}{16(1+\frac{1}{\gamma_{ps}})m\tau^*_m + 80\theta} \leq \dfrac{1}{80}$, we have

$$\left(1 - \frac{\theta}{(16(1+\frac{1}{\gamma_{ps}})m\tau^*_m + 80\theta)2\ln(2)}\right)^2 - \frac{1}{2} \geq 0,$$

and

$$\mathbb{P}\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\hat{k}}\right) - \mathbb{L}(g^*) - (1+\theta)\left(\mathbb{L}\left(\left(\hat{g}_1^n\right)_{\tilde{k}}\right) - \mathbb{L}(g^*)\right) > \varepsilon\right) \leq$$
$$\left(2\exp\left(\frac{\ln(2)}{t_{mix}}\right) + N\right)\exp\left(-\frac{1}{4t_{mix}(1+\theta)}\frac{\theta\gamma_{ps}\varepsilon m}{16(1+\frac{1}{\gamma_{ps}})m\tau^*_m + 80\theta}\right). \tag{A.10}$$

■

Annexe B

Appendices du chapitre 3

B.1 Preuves du Corollaire 3.3

Démonstration.

If $(\mathbf{H}_1^\infty)$, $(\mathbf{H}_2^\infty)$ and $(\mathbf{H}_3)$ are verified,

$$\begin{aligned}
&\mathbb{E}[\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|)] = \int_{t=0}^{\infty} \mathbb{P}\left(\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|) \geq t\right) dt \\
&= \int_{t=0}^{\exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)} \mathbb{P}\left(\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|) \geq t\right) dt \\
&+ \int_{t=\exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)}^{\infty} \mathbb{P}\left(\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|) \geq t\right) dt.
\end{aligned}$$

From one side,

$$\int_{t=0}^{\exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)} \mathbb{P}\left(\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|) \geq t\right) dt \leq \exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right).$$

From the other side,

$$\begin{aligned}
&\int_{t=\exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)}^{\infty} \mathbb{P}\left(\exp(s|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})|) \geq t\right) dt \\
&= \int_{t=\exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)}^{\infty} \mathbb{P}\left(|\mathbb{L}(\hat{f}) - \widehat{\mathrm{L}}_n(\hat{f})| \geq \frac{\ln(t)}{s}\right) dt \\
&= \int_{t=\exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)}^{\infty} \exp\left(\frac{-2n^2\left(\frac{\ln(t)}{s} - \frac{2n_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right) dt \\
&\leq \exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}\right)\int_{t=1}^{\infty} 2\exp\left(\frac{-2n^2\ln(t)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right) dt.
\end{aligned}$$

Then,

$$\begin{aligned}
&\int_{t=1}^{\infty} 2\exp\left(\frac{-2n^2\ln(t)^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right) dt \\
&= 2\int_{t=0}^{\infty} \exp\left(\frac{-2n^2t^2}{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)} + t\right) dt \\
&\leq \frac{n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s}{n}\sqrt{2\pi\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\exp\left(\frac{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s\right)^2\left(1 + \mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}{8n^2}\right).
\end{aligned}$$

It yields,

$$\mathbb{E}[\exp(s|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|)] \leq \exp\left(\frac{2sn_{\bar{\mathrm{B}}}\mathbb{V}_\infty}{n}+\frac{\left(n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s\right)^2\left(1+\mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}{8n^2}\right)$$
$$\times\left(1+\frac{n_{\bar{\mathrm{B}}}\mathbb{V}_\infty s}{n}\sqrt{2\pi\left(1+\mathrm{A}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3\kappa!^2\lceil\ln(n)\rceil^\kappa n\right)}\right).$$

If $(\mathbf{H}_1^{\mathbf{m}})$, $(\mathbf{H}_2^{\mathbf{m}})$, $(\mathbf{H_3})$ and $(\mathbf{H_4})$ are verified. The demonstration is very similar to the previous point. Indeed, we have,

$$\begin{aligned}
&\mathbb{E}[\exp(s|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|)]\\
&\leq \exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)+\int_{t=\exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)}^{\exp(\mathrm{M}s)}\mathbb{P}\left(|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|\geq\frac{\ln(t)}{s}\right)dt\\
&\leq \exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)+\int_{t=\exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)}^{\exp(\mathrm{M}s)}2\exp\left(\frac{-2n^{2-\frac{4}{m}}\left(\frac{\ln(t)}{2s}-\frac{\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)^2}{\left(\mathrm{H}n_{\bar{\mathrm{B}}}\right)^2\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right)+\frac{\rho^m\mathrm{L}_2(n)}{n}dt.
\end{aligned}$$

And,

$$\begin{aligned}
&\int_{t=\exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)}^{\exp(\mathrm{M}s)}2\exp\left(\frac{-2n^{2-\frac{4}{m}}\left(\frac{\ln(t)}{2s}-\frac{\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)^2}{\left(\mathrm{H}n_{\bar{\mathrm{B}}}\right)^2\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right)dt\\
&=\exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)\int_{t=1}^{\infty}2\exp\left(\frac{-n^{2-\frac{4}{m}}\ln(t)^2}{2\left(\mathrm{H}n_{\bar{\mathrm{B}}}s\right)^2\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right)dt\\
&=2\exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)\int_{t=0}^{\infty}\exp\left(\frac{-n^{2-\frac{4}{m}}t^2}{2\left(\mathrm{H}n_{\bar{\mathrm{B}}}s\right)^2\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}+t\right)dt\\
&=2\exp\left(\frac{2s\mathrm{L}_1(n)}{n^{1-\frac{2}{m}}}\right)\frac{\mathrm{H}n_{\bar{\mathrm{B}}}s}{n^{1-\frac{2}{m}}}\sqrt{2\pi\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\\
&\times\exp\left(\frac{\left(\mathrm{H}n_{\bar{\mathrm{B}}}s\right)^2\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}{2n^{2-\frac{4}{m}}}\right)\\
&\leq 2\exp\left(\frac{1}{2n^{1-\frac{4}{m}}}\left(4s\mathrm{L}_1(n)+\left(\mathrm{H}n_{\bar{\mathrm{B}}}s\right)^2\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)\right)\right)\\
&\times\frac{\mathrm{H}n_{\bar{\mathrm{B}}}s}{n^{1-\frac{2}{m}}}\sqrt{2\pi\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}.
\end{aligned}$$

Consequently,

$$\mathbb{E}[\exp(s|\mathbb{L}(\hat{f})-\widehat{\mathrm{L}}_n(\hat{f})|)] \leq \exp\left(\frac{1}{2n^{1-\frac{4}{m}}}\left(4s\mathrm{L}(n)+\left(\mathrm{H}n_{\bar{\mathrm{B}}}s\right)^2\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)\right)\right)$$
$$\times\left(1+\frac{\exp\left(\mathrm{M}s\right)\rho^m\mathrm{L}_2(n)}{n}+\frac{2\mathrm{H}n_{\bar{\mathrm{B}}}s}{n^{1-\frac{2}{m}}}\sqrt{2\pi\left(1+\mathrm{E}n_{\bar{\mathrm{B}}}n_{\mathrm{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^\kappa n\right)}\right).$$

□

B.2 Preuve du Lemme 3.2

Démonstration. Let's prove by induction that for each i in $[0,d]$,

$$\forall t\in\mathbb{Z}^\kappa,\forall d\in\mathbb{N},\|\mathrm{X}_t-\tilde{\mathrm{X}}_t^{[i]}\|_m\leq\rho^i\mathbb{V}_m. \tag{B.1}$$

For $i=0$, $\tilde{X}_t^{[0]}=\bar{X}$ and $\bar{X}$ is drawn from the law μ_X. Therefore, using hypothesis $(\mathbf{H_2^m})$, we get

$$\forall t\in\mathbb{Z}^\kappa, \|X_t-\tilde{X}_t^{[0]}\|_m\le \mathbb{V}_m.$$

Thus Equation (B.1) is verified for $i=0$.
Moreover, if we suppose that (B.1) is verified for i. For each $t\in\mathbb{Z}^\kappa$, it holds

$$\begin{aligned}\|X_t-\tilde{X}_t^{[i+1]}\|_m &\le \|F((X_{t+s})_{s\in\mathscr{B}},\varepsilon_t)-F\left((\tilde{X}_{t+s}^{[i]})_{s\in\mathscr{B}},\varepsilon_t\right)\|_m\\ &\le \sum_{s\in\mathscr{B}}\lambda_s\|X_{t+s}-\tilde{X}_{t+s}^{[i]}\|_m\le\sum_{s\in\mathscr{B}}\lambda_s\rho^i\mathbb{V}_m\le\rho^{i+1}\mathbb{V}_m.\end{aligned}$$

Consequently, by induction, Equation (B.1) is verified for each i in $[0,d]$. Finally, setting $i=d$ yields Lemma 3.2. □

B.3 Preuve du Lemme 3.6

Démonstration.

We first show that $\|\tilde{S}_{\mathscr{I}}^{[d]}-\tilde{S}_{\mathscr{I}}^{[d]\prime}\|_m\le nn_{\bar{B}}\rho^d\mathbb{V}_m$.

$$\begin{aligned}\|\tilde{S}_{\mathscr{I}}^{[d]}-\tilde{S}_{\mathscr{I}}^{[d]\prime}\|_m &= \|\sum_{t\in\mathscr{I}}\Phi((\tilde{X}_{t+s}^{[d]})_{s\in\bar{\mathscr{B}}})-\sum_{t\in\mathscr{I}}\Phi((\tilde{X}_{t+s}^{[d]\prime})_{s\in\bar{\mathscr{B}}})\|_m\\ &\le\sum_{t\in\mathscr{I}}\|\Phi((\tilde{X}_{t+s}^{[d]})_{s\in\bar{\mathscr{B}}})-\Phi((\tilde{X}_{t+s}^{[d]\prime})_{s\in\bar{\mathscr{B}}})\|_m\\ &\le\sum_{t\in\mathscr{I}}\sum_{s\in\bar{\mathscr{B}}}\|\tilde{X}_{t+s}^{[d]}-\tilde{X}_{t+s}^{[d]\prime}\|_m\\ &\le nn_{\bar{B}}\rho^d\mathbb{V}_m \text{ using Lemma 3.5.}\end{aligned}$$

We then bound $\|\tilde{S}_{\mathscr{I}}^{[d]}-\tilde{S_{\mathscr{I}}}^{[d,i]}\|_m$. For all t and i $\mathbb{Z}^\kappa$ we have the following properties

$$\tilde{X}_t^{[d]}\neq\tilde{X}_t^{[d,i]}\iff i\in\mathcal{V}(d\delta,t)\iff t\in\mathcal{V}(d\delta,i).$$

$$\begin{aligned}\|\tilde{S_{\mathscr{I}}}^{[d,i]}-\tilde{S}_{\mathscr{I}}^{[d]}\|_m &= \|\sum_{t\in\mathscr{I}}\Phi((\tilde{X}_{t+s}^{[d]})_{s\in\bar{\mathscr{B}}})-\sum_{t\in\mathscr{I}}\Phi((\tilde{X}_{t+s}^{[d,i]})_{s\in\bar{\mathscr{B}}})\|_m\\ &\le\sum_{t\in\mathscr{I}}\sum_{s\in\bar{\mathscr{B}}}\|\tilde{X}_{t+s}^{[d]}-\tilde{X}_{t+s}^{[d,i]}\|_m\\ &\le\sum_{t\in\mathbb{Z}^\kappa}\sum_{s\in\bar{\mathscr{B}}+t}\|\tilde{X}_s^{[d]}-\tilde{X}_s^{[d,i]}\|_m.\end{aligned}$$

Only random variables $\tilde{X}_t^{[d,i]}$ with $t\in\mathcal{V}(d\delta,i)$ are impacted by the replacement of ε_i by ε_i'. Thus, at worst, each $t\in\mathcal{V}(d\delta,i)$ appears $n_{\bar{B}}$ times in the sum. Therefore, we get

$$\|\tilde{S_{\mathscr{I}}}^{[d,i]}-\tilde{S}_{\mathscr{I}}^{[d]}\|_m\le\sum_{t\in\mathcal{V}(d\delta,i)}n_{\bar{B}}\|\tilde{X}_t^{[d,i]}-\tilde{X}_t^{[d]}\|_m.$$

Moreover, for all t in $\mathbb{Z}^\kappa$

$$\mathcal{V}(d\delta,t)=\bigcup_{c=0}^{d}\mathcal{V}(c\delta,t)\setminus\mathcal{V}((c-1)\delta,t)\quad\text{with}\quad\mathcal{V}(0,t)=t\quad\text{and}\quad\mathcal{V}(-1,t)=\emptyset.$$

And by definition of $\mathcal{V}(d\delta,t)$ (see Section 3.2.1), it holds

$$\forall r_1\neq r_2,(\mathcal{V}(r_1\delta,t)\setminus\mathcal{V}((r_1-1)\delta,t))\bigcap(\mathcal{V}(r_2\delta,t)\setminus\mathcal{V}((r_2-1)\delta,ts))=\emptyset.$$

Therefore, $\bigcup_{c=0}^{d} \mathcal{V}(c\delta,t)\backslash\mathcal{V}(c-1,t)$ is a partition of the set $\mathcal{V}(d\delta,t)$ and we can rewrite the previous inequality as

$$\|\tilde{\mathrm{S}}_{\mathscr{I}}^{[d,i]} - \tilde{\mathrm{S}}_{\mathscr{I}}^{[d]}\|_m \leq n_{\tilde{\mathrm{B}}} \sum_{c=0}^{d} \left(\sum_{s\in\mathcal{V}(c\delta,i)\backslash\mathcal{V}((c-1)\delta,i)} \|\tilde{\mathrm{X}}_s^{[d,i]} - \tilde{\mathrm{X}}_s^{[d]}\|_m \right).$$

Using Lemma 3.5. We get

$$\|\tilde{\mathrm{S}}_{\mathscr{I}}^{[d,i]} - \tilde{\mathrm{S}}_{\mathscr{I}}^{[d]}\|_m \leq n_{\tilde{\mathrm{B}}} \sum_{c=0}^{d} \mathrm{Card}(\mathcal{V}(c\delta,i)\backslash\mathcal{V}((c-1)\delta,i))\rho^c \mathbb{V}_m.$$

$\forall c \in \mathbb{N}, \mathcal{V}(c\delta,i)$ is a $\kappa-$orthotope and

$$\mathrm{Card}(\mathcal{V}(c\delta,i)) = \prod_{j=1}^{\kappa} \left(2c\delta_j + 1\right).$$

We recall that $n_{\mathscr{B}} = \prod_{j=1}^{\kappa} \left(2\delta_j + 1\right) = \mathrm{Card}(\mathscr{B}) + 1$. Then, for all $c > 1$,

$$\begin{aligned}
\mathrm{Card}(\mathcal{V}(c\delta,i)\backslash\mathcal{V}((c-1)\delta,i)) &= \mathrm{Card}(\mathcal{V}(c\delta,i)) - \mathrm{Card}(\mathcal{V}((c-1)\delta,i)) \\
&= \prod_{j=1}^{\kappa} \left(2c\delta_j + 1\right) - \prod_{j=1}^{\kappa} \left(2(c-1)\delta_j + 1\right) \\
&\leq c^{\kappa} \prod_{j=1}^{\kappa} \left(2\delta_j + \frac{1}{c}\right) - (c-1)^{\kappa} \prod_{j=1}^{\kappa} \left(2\delta_j + \frac{1}{c-1}\right) \\
&\leq \prod_{j=1}^{\kappa} \left(2\delta_j + \frac{1}{c-1}\right) \left(c^{\kappa} - (c-1)^{\kappa}\right) \\
&\leq \prod_{j=1}^{\kappa} \left(2\delta_j + 1\right) \left(c^{\kappa} - (c-1)^{\kappa}\right) \\
&\leq n_{\mathscr{B}} \left(c^{\kappa} - (c-1)^{\kappa}\right) \leq n_{\mathscr{B}} \kappa c^{\kappa-1}.
\end{aligned}$$

Eventually, we get $\|\tilde{\mathrm{S}}_{\mathscr{I}}^{[d,i]} - \tilde{\mathrm{S}}_{\mathscr{I}}^{[d]}\|_m \leq n_{\tilde{\mathrm{B}}} \mathbb{V}_m \left(1 + \sum_{c=1}^{d} n_{\mathscr{B}} \kappa c^{\kappa-1} \rho^c\right)$. □

B.4 Preuve du Lemme 3.7

Démonstration. Let $p \in \mathbb{N}$ and $(a,b) \in \mathbb{R}^2$.
We first compute $\mathrm{I}_p = \int_a^b t^p \rho^t \, dt$.

$$\mathrm{I}_p = \int_a^b t^p \rho^t \, dt = \left[t^p \frac{\rho^t}{\ln(\rho)} \right]_a^b + \frac{p}{\ln(\rho^{-1})} \int_a^b t^{p-1} \rho^t \, dt = \frac{a^p \rho^a - b^p \rho^b}{\ln(\rho^{-1})} + \frac{p}{\ln(\rho^{-1})} \mathrm{I}_{p-1}.$$

And

$$\mathrm{I}_0 = \int_a^b \rho^t \, dt = \frac{\rho^a - \rho^b}{\ln(\rho^{-1})}.$$

By induction, we get

$$\mathrm{I}_p = \sum_{i=0}^{p} \frac{a^i \rho^a - b^i \rho^b}{\ln(\rho^{-1})^{p-i+1}} \times \frac{p!}{i!} = \frac{p!}{\ln(\rho^{-1})^{p+1}} \sum_{i=0}^{p} \frac{\left(a^i \rho^a - b^i \rho^b\right) \ln(\rho^{-1})^i}{i!}. \tag{B.2}$$

Let $f(t) = t^p \rho^t$; we have $f'(t) = t^{p-1} \rho^t \left(p - t\ln(\rho^{-1})\right)$. Then $f'(t) = 0 \iff t = \frac{p}{\ln(\rho^{-1})}$. Thus f is increasing on $[0, \frac{p}{\ln(\rho^{-1})}]$ and decreasing on $[\frac{p}{\ln(\rho^{-1})}, +\infty]$.
Applying this to our case,

— if $d < \lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor$. Then : $\sum_{c=1}^{d} c^{\kappa-1}\rho^c \leq \int_0^{d+1} t^{\kappa-1}\rho^t dt$.

— if $d = \lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor$, Then $\sum_{c=1}^{d} c^{\kappa-1}\rho^c \leq \int_0^{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor} t^{\kappa-1}\rho^t dt + \left(\frac{\kappa-1}{\ln(\rho-1)e}\right)^{\kappa-1}$.

— if $d > \lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor$. Then $\sum_{c=1}^{d} c^{\kappa-1}\rho^c \leq \int_0^{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor} t^{\kappa-1}\rho^t dt + \left(\frac{\kappa-1}{\ln(\rho-1)e}\right)^{\kappa-1} + \int_{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor}^{d} t^{\kappa-1}\rho^t dt$.

Using Equation (B.2), we get

— $\int_0^{d+1} t^{\kappa-1}\rho^t dt \leq \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(1 - \rho^{d+1}\sum_{i=0}^{\kappa-1}\frac{\left((d+1)\ln(\rho^{-1})\right)^i}{i!}\right)$.

— $\int_{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor}^{d} t^{\kappa-1}\rho^t dt \leq \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(\rho^{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor}\sum_{i=0}^{\kappa-1}\frac{\left(\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor \ln(\rho^{-1})\right)^i}{i!} - \rho^d\sum_{i=0}^{\kappa-1}\frac{\left(d\ln(\rho^{-1})\right)^i}{i!}\right)$.

We recall that,

$$\|\tilde{S}_{\mathscr{J}}^{[d,i]} - \tilde{S}_{\mathscr{J}}^{[d]}\|_m \leq n_{\bar{B}}\mathbb{V}_m\left(1 + n_B\kappa\sum_{c=1}^{d} c^{\kappa-1}\rho^c\right) \leq n_{\bar{B}}n_B\mathbb{V}_m\kappa\left(\frac{1}{n_B\kappa} + \sum_{c=1}^{d} c^{\kappa-1}\rho^c\right).$$

We now define the function Υ.

$\Upsilon : \mathbb{N} \mapsto \mathbb{R}$

$$\Upsilon(d) = \begin{cases} \kappa\left(\frac{1}{n_B\kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(1 - \rho^{d+1}\sum_{i=0}^{\kappa-1}\frac{\left((d+1)\ln(\rho^{-1})\right)^i}{i!}\right)\right), & \text{if } d < \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \\ \kappa\left(\frac{1}{n_B\kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(1 - e^{-(\kappa-1)}\sum_{i=0}^{\kappa-1}\frac{\left(\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor \ln(\rho^{-1})\right)^i}{i!}\right) + \left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}\right), & \text{if } d = \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \\ \kappa\left(\frac{1}{n_B\kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa}\left(1 - \rho^d\sum_{i=0}^{\kappa-1}\frac{\left(d\ln(\rho^{-1})\right)^i}{i!}\right) + \left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}\right), & \text{if } d > \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \end{cases}$$

Consequently, we proved that,

$$\|\tilde{S}_{\mathscr{J}}^{[d,i]} - \tilde{S}_{\mathscr{J}}^{[d]}\|_m \leq n_{\bar{B}}n_B\mathbb{V}_m\Upsilon(d).$$

It is important to note that we can easily verify that the function Υ is increasing and bounded by its limit when $d \to \infty$. It holds

$$\begin{aligned} \forall d \in \mathbb{N}, \Upsilon(d) \leq \lim_{d\to\infty}\Upsilon(d) &= \kappa\left(\frac{1}{n_B\kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa} + \left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}\right) \\ &\leq \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa\left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}. \end{aligned}$$

Thereafter, we will denote $\upsilon = \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa\left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}$. Moreover, we emphasize that using Stirling formula, we can proved that $\upsilon \underset{\kappa\to\infty}{\sim} \frac{\kappa!}{\ln(\rho^{-1})^\kappa}$.

□

B.5 Preuve du Lemme 3.9

Démonstration. First, we establish a A-difference bound (Definition 3.9). Let $t_1, t_2 > 0$, using Lemma 3.6, 3.7 and Markov's inequality, it holds

$$\forall t_1 > 0, \mathbb{P}(|\tilde{S}_{\mathscr{J}}^{[d]} - \tilde{S}_{\mathscr{J}}^{[d]\prime}| \geq t_1) \leq \left(\frac{n n_{\bar{B}}\rho^d\mathbb{V}_m}{t_1}\right)^m.$$

Then, for all $i \in \bigcup_{t\in\mathscr{J}} \mathcal{V}(d\bar{\delta}, t)$, for all $t_2 > 0, \mathbb{P}(|\tilde{S}_{\mathscr{J}}^{[d]} - \tilde{S_{\mathscr{J}}}^{[d,i]}| \geq t_2) \leq \left(\frac{n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d)}{t_2}\right)^m$.

We define the event $A = \bigcup_{i\in\bigcup_{t\in\mathscr{J}}\mathcal{V}(d\bar{\delta},t)} \left(|\tilde{S}_{\mathscr{J}}^{[d]} - \tilde{S_{\mathscr{J}}}^{[d,i]}| > t_2\right) \bigcup \left(|\tilde{S}_{\mathscr{J}}^{[d]} - \tilde{S}_{\mathscr{J}}^{[d]\prime}| > t_1\right)$ and $p = 1 - \mathbb{P}(A)$.

By union bound, it holds $p \leq \left(\frac{n n_{\bar{B}} \rho^d \mathbb{V}_m}{t_1}\right)^m + N_2 \left(\frac{n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d)}{t_2}\right)^m$. We also define $\bar{c} = t_1 + N_2 t_2$.

Therefore, applying Theorem 2.1 from [Com15], we get

$$\forall \varepsilon > 0, \mathbb{P}\left(|\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|A\right]| \geq \varepsilon\right) \leq 2\left(p + \exp\left(\frac{-2(\varepsilon - p\bar{c})^2}{t_1^2 + N_2 t_2^2}\right)\right).$$

Moreover

$$\mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right] = \mathbb{P}(A)\mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|A\right] + \mathbb{P}(\bar{A})\mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|\bar{A}\right].$$

Then

$$\begin{aligned}|\mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right] - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|A\right]| &= (1 - \mathbb{P}(A))\,|\mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|\bar{A}\right] - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|A\right]| \\ &\leq p 2 n M \text{ with } M = \|\Phi\|_\infty \text{ (see Hypothesis } (\mathbf{H_4})).\end{aligned}$$

Eventually, we get

$$\begin{aligned}\forall \varepsilon > 2npM, \mathbb{P}\left(|\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right]| \geq \varepsilon\right) &= \mathbb{P}\left(|\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|A\right] + \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|A\right] - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}\right]| \geq \varepsilon\right) \\ &\leq \mathbb{P}\left(|\tilde{S}_{\mathscr{J}}^{[d]} - \mathbb{E}\left[\tilde{S}_{\mathscr{J}}^{[d]}|A\right]| \geq \varepsilon - 2npM\right) \\ &\leq 2\left(p + \exp\left(\frac{-2(\varepsilon - (p\bar{c} + 2npM))^2}{t_1^2 + N_2 t_2^2}\right)\right).\end{aligned}$$

□

B.6 Preuve du Théorème 3.6

Démonstration. From Lemma 3.3, we get $N_2 \leq N_1 n_d \leq n_{\bar{B}} n_B d^\kappa n$.

We set $\tilde{d} = \frac{(1-\frac{1}{m})\ln(n)}{\ln(\rho^{-1})}$, $d = \lceil \tilde{d} \rceil$, $t_1 = \frac{n^{1+\frac{1}{m}} \rho^{\tilde{d}} n_{\bar{B}} \mathbb{V}_m}{\rho (n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}}}$ and $t_2 = \frac{n^{\frac{2}{m}} n_B n_{\bar{B}} \Upsilon(d) \mathbb{V}_m}{\rho}$. Then

$$\begin{aligned}p &\leq \left(\frac{n n_{\bar{B}} \rho^d \mathbb{V}_m}{t_1}\right)^m + N_2 \left(\frac{n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d)}{t_2}\right)^m \\ &\leq \left(\frac{n n_{\bar{B}} \rho^{\tilde{d}} \mathbb{V}_m}{t_1}\right)^m + N_2 \left(\frac{n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d)}{t_2}\right)^m \\ &\leq \frac{\rho^m n_{\bar{B}} n_B d^\kappa}{n} + n_{\bar{B}} n_B d^\kappa n \frac{\rho^m}{n^2} \\ &\leq \frac{2 n_{\bar{B}} n_B d^\kappa \rho^m}{n}.\end{aligned}$$

We also note that $\rho^{\tilde{d}} = \rho^{\tilde{d}} = \exp(-\ln(\rho^{-1}) K_\rho(m) \ln(n)) = \exp(-\left(1 - \frac{1}{m}\right)\ln(n)) = n^{-\left(1-\frac{1}{m}\right)}$.

On the other hand,

$$\begin{aligned}t_1^2 + N_2 t_2^2 &\leq \left(\frac{n^{1+\frac{1}{m}} \rho^{\tilde{d}} n_{\bar{B}} \mathbb{V}_m}{\rho (n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}}}\right)^2 + n_{\bar{B}} n_B d^\kappa n \left(\frac{n^{\frac{2}{m}} n_B n_{\bar{B}} \Upsilon(d) \mathbb{V}_m}{\rho}\right)^2 \\ &= \left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho}\right)^2 \left(\frac{1}{(n_{\bar{B}} n_B d^\kappa)^{\frac{2}{m}}} + n n_{\bar{B}} n_B^3 \Upsilon(d)^2 d^\kappa\right).\end{aligned}$$

And

$$\begin{aligned}\bar{c} = t_1 + \mathrm{N}_2 t_2 &\leq \frac{n^{1+\frac{1}{m}}\rho^{\tilde{d}} n_{\bar{\mathrm{B}}} \mathbb{V}_m}{\rho (n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{1}{m}}} + n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa} n \left(\frac{n^{\frac{2}{m}} n_{\mathrm{B}} n_{\bar{\mathrm{B}}} \Upsilon(d) \mathbb{V}_m}{\rho} \right) \\ &\leq \left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right) \left(\frac{1}{(n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{1}{m}}} + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^2 \Upsilon(d) d^{\kappa} n \right).\end{aligned}$$

Therefore,

$$\begin{aligned}2npM + p\bar{c} &\leq \frac{2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa} \rho^m}{n} \left(\left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right) \left(\frac{1}{(n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{1}{m}}} + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^2 \Upsilon(d) d^{\kappa} n \right) + 2nM \right) \\ &\leq 2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa} \rho^m \left(\left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right) \left(\frac{1}{(n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{1}{m}} n} + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^2 \Upsilon(d) d^{\kappa} \right) + 2M \right) \\ &\leq 2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^m n^{\frac{2}{m}} \left(\left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m}{\rho} \right) \left(\frac{1}{d^{\kappa} (n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{1}{m}} n} + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^2 \Upsilon(d) \right) + \frac{2M}{d^{\kappa} n^{\frac{2}{m}}} \right) \\ &\leq 2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^m n^{\frac{2}{m}} \mathrm{L}(n).\end{aligned}$$

With $\mathrm{L}(n) = \left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m}{\rho} \right) \left(\frac{1}{d^{\kappa} (n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{1}{m}} n} + n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^2 \Upsilon(d) \right) + \frac{2M}{d^{\kappa} n^{\frac{2}{m}}}$.
Consequently using $p \leq \frac{2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa} \rho^m}{n}$. It holds $\forall \varepsilon > 4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^m n^{\frac{2}{m}} \mathrm{L}(n)$,

$$\exp\left(\frac{-2(\frac{\varepsilon}{2} - (p\bar{c} + 2npM))^2}{t_1^2 + \mathrm{N}_2 t_2^2} \right) \leq \exp\left(\frac{-2\left(\frac{\varepsilon}{2} - 2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^m n^{\frac{2}{m}} \mathrm{L}(n) \right)^2}{\left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right)^2 \left(\frac{1}{(n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{2}{m}}} + n n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^{\kappa} \right)} \right).$$

In particular; $\forall \varepsilon > 4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} \mathrm{L}(n)$,

$$\exp\left(\frac{-2(\frac{\varepsilon}{2} - (p\bar{c} + 2npM))^2}{t_1^2 + \mathrm{N}_2 t_2^2} \right) \leq \exp\left(\frac{-2\left(\frac{\varepsilon}{2} - 2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} \mathrm{L}(n) \right)^2}{\left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right)^2 \left(\frac{1}{(n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{2}{m}}} + n n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^{\kappa} \right)} \right).$$

Finally,

$$\begin{aligned}\forall \varepsilon > 4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} \mathrm{L}(n), \left(\frac{2n n_{\bar{\mathrm{B}}} \rho^d \mathbb{V}_m}{\varepsilon} \right)^m &\leq \left(\frac{2n n_{\bar{\mathrm{B}}} \rho^{\tilde{d}} \mathbb{V}_m}{4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} \mathrm{L}(n)} \right)^m \\ &\leq \frac{\rho^m}{n} \left(\frac{\mathbb{V}_m}{2n_{\mathrm{B}} d^{2\kappa} \mathrm{L}(n)} \right)^m.\end{aligned}$$

Using Equation (3.24), for $\varepsilon > 4n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} \mathrm{L}(n)$, we get

$$\begin{aligned}\mathbb{P}\left(|\mathrm{S}_{\mathscr{I}} - \mathbb{E}[\mathrm{S}_{\mathscr{I}}]| \geq \varepsilon \right) \leq & 2\exp\left(\frac{-2\left(\frac{\varepsilon}{2} - 2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} \mathrm{L}(n) \right)^2}{\left(\frac{n_{\bar{\mathrm{B}}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right)^2 \left(\frac{1}{(n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa})^{\frac{2}{m}}} + n n_{\bar{\mathrm{B}}} n_{\mathrm{B}}^3 \Upsilon(d)^2 d^{\kappa} \right)} \right) \\ &+ \frac{2n_{\bar{\mathrm{B}}} n_{\mathrm{B}} d^{\kappa} \rho^m}{n} + \frac{\rho^m}{n} \left(\frac{\mathbb{V}_m}{2n_{\mathrm{B}} d^{2\kappa} \mathrm{L}(n)} \right)^m.\end{aligned}$$

□

Annexe C

Appendices du chapitre 6 :Borne de moment pour une distribution Poisson

Lemme C.1. *Let* X *be a random variable following a Poisson distribution of parameter* λ. *Denote* $M = \max(1, \lambda e)$, *then :*

$$\mathbb{E}\left[|X|^k\right] \leq k! M^k.$$

Démonstration. The moment generating function of the X is $g(x) = \exp(\lambda(e^x - 1))$. We denote by $m_k = g^{(k)}(0)$ the k-th moment of the distribution. The first derivative of g satisfies $g'(x) = \lambda \exp(x) g(x)$. Using Leibniz formula, we have :

$$g^{(k+1)}(x) = \lambda \sum_{i=0}^{k} \binom{n}{i} g^{(i)}(x) \exp(x).$$

For all k, we have the following recurrent relation

$$m_{k+1} \leq \lambda \sum_{i=0}^{k} \binom{n}{i} m_i.$$

We will prove the hypothesis $H_k : m_k \leq k! M^k$ by induction. We have $m_0 = 1$ and $m_1 = \lambda$, so H_0 and H_1 are verified. For $k > 1$, if we suppose (H_i) verified for all $i \leq K$:

$$\begin{aligned} m_{k+1} &\leq \lambda \sum_{i=0}^{k} \binom{k}{i} i! M^i \leq \lambda \sum_{i=0}^{k} \frac{k!}{(k-i)!} M^i \leq \lambda M^k k! \sum_{i=0}^{k} \frac{1}{i!} M^{-i} \\ &\leq \lambda M^k (k+1)! e^{\frac{1}{M}} \end{aligned}$$

As $M \geq 1 : m_{k+1} \leq \mu e M^k (k+1)! \leq M^{k+1}(k+1)!$. □

[Ahl21] gives slightly better moment bound for Poisson distribution. However, this article has not been published yet.