Learning under Dependence for Aggregation of Estimators andClassification, with Applications to DNA Analysis

Xiaoyin Li

École Doctorale EM2C

THÈSE DE DOCTORAT

Discipline : Mathématiques

présentée par

Xiaoyin LI

Learning under Dependence for Aggregation of Estimators and Classification, with Applications to DNA Analysis

---

dirigée par Paul DOUKHAN

Soutenue le 23 octobre devant le jury composé de :

Pierre ALQUIER	ENSAE	Examinateur
Gabriel LANG	AgroParisTech	Examinateur
Eva LÖCHERBACH	Univ. de Cergy-Pontoise	Examinateur
Jean-Marc BARDET	Univ. Paris I	Examinateur
Donatas SURGAILIS	Vilnius University	Examinateur
Pascal MASSART	Univ. Paris-Sud	Rapporteur
Konstantinos FOKIANOS	University of Cyprus	Rapporteur
Paul DOUKHAN	Univ. de Cergy-Pontoise	Directeur

Laboratoire AGM
2 Avenue Adolphe-Chauvin
95 302 Cergy-Pontoise

École doctorale EM2C $N°405$
33 Boulevard du Port
95 011 Cergy-Pontoise

Remerciements

Mes remerciements sincères et toute ma gratitude vont au Professeur Paul Doukhan, mon directeur de thèse. Il a su me proposer un sujet motivant qui a avivé mon goût pour la recherche; Il m'a permis de participer à plusieurs conférences et écoles d'été afin de m'ouvrir à d'autres points vue et accéder à de nouvelles connaissances. Je porte à son égard le plus grand estime pour ce qu'il a toujours été : ingénieux et rigoureux à toute épreuve, dans l'humilité spontanée et la bienveillance perspicace.

Professeur Pascal Massart et Professeur Konstantinos Fokianos ont accepté d'être les rapporteurs de cette thèse, et je les en remercie, de même que pour leur participation au Jury. Ils ont également contribué par leurs nombreuses remarques et suggestions à améliorer la qualité de ce mémoire, et je leur en suis très reconnaissant.

Merci également aux autres membres du jury qui ont accepté de juger ce travail : Pierre Alquier, Gabriel Lang, Eva Löcherbach, Donatas Surgailis, Jean-Marc Bardet. Je leur adresse mes profonds remerciements.

Je tiens à remercier tous les membres du Laboratoire AGM au sein du quel j'ai effectué me thèse. Merci pour le climat sympathique dans lequel ils m'ont permis de travailler. Les nombreuses discussions que j'ai pu avoir avec chacun m'ont beaucoup apporté. Merci donc à Vladimir Georgescu, Lysianne Haril, Lorenzo Pittau, Marie Carette, Linda Isone, Thomas Ballesteros. Je voudrais exprimer particulièrement toute mon amitié à Chao Wang, Xiaoyi Chen, José Gomez pour leur gentillesse, leur compétence et leur humour.

J'adresse toute ma gratitude à tous mes amis et à toutes les personnes qui m'ont aidé dans la réalisation de ce travail. Je remercie Matthieu Cornec et Aurélien D'isanto pour m'avoir offert les données et le modèle dans l'application de la prédiction quantile du PIB français. Merci à Olivier Wintenbeger avec qui je partage un article. Un grand merci, aux personnes qui m'ont conseillé et avec qui j'espère collaborer dans le futur comme Johan Segers, Sylvain Arlot, Gille Scoltz.

Mention spéciale à Louise Da Silveira et Diane Baratier, qui m'ont toujours expliqué patiemment les mots français et m'ont beaucoup aidé dans la vie quotidienne. Je vos remercie profondément.

Finalement, les mots les plus simples étant les plus forts, j'adresse toute mon

affection à ma famille en chine. Malgré mon éloignement depuis de nombreuses années, leurs intelligence, leur confiance, leur tendresse, leur amour me portent et me guident tous les jours. Sans leur soutien et leurs encouragements, je n'aurais pas pu surmonter les difficultés de tout ordre que j'ai pu rencontrer. Merci pour avoir fait de moi ce que je suis aujourd'hui. Je vous aime.

Résumé

Dans cette thèse, nous donnons une introduction systématique à la condition dépendance faible, introduit par Doukhan and Louhichi (1999) , qui est plus générale que les cadres classiques de mélange ou de séquences associées. La notion est suffisamment large pour inclure des modèles standards tels que les modèles stables de Markov , les modèles bilinéaires , et plus généralement , les schémas de Bernoulli. Dans certains cas, aucunes des propriétés de mélangeant ne peut s'attendre sans hypothèse de régularité supplémentaire sur la distribution innovations pour lesquelles une condition de dépendance faible peut être facilement dérivée. Nous étudions la relation entre dépendance faible et mélangeant pour les processus de valeurs discrètes. Nous montrons que la dépendance faible implique des conditions de mélangeant sous des hypothèses naturelles. Les résultats se spécialisent au cas des processus Markovian. Plusieurs exemples de processus à valeur entier sont examinés et leurs propriétés de dépendance faibles sont étudiés à l'aide d'une contraction principale.

Dans la deuxième partie, nous établissons des vitesses de convergences en apprentissage statistique pour les prédictions d'une série chronologique. En utilisant l'approche PAC- bayésienne, les vitesses lentes de convergence $\sqrt{d/n}$ pour l'estimateur de Gibbs sous la perte absolue ont été donnés dans un travail précédent Alquier and Wintenberger (2012), où n est la taille de l'échantillon et d la dimension de l'ensemble des prédicteurs. Sous les mêmes conditions de dépendance faible, nous étendons ce résultat à une fonction de perte Lipschitz convexe. Nous identifions également une condition sur l'espace des paramètres qui assure des vitesses similaires pour la procédure classique de l'ERM pénalisé. Nous appliquons cette méthode pour la prédiction quantile du PIB français. Dans des conditions supplémentaires sur les fonctions de perte (satisfaites par la fonction de perte quadratique) et pour les processus uniformément mélangeant, nous montrons que l'estimateur de Gibbs atteint effectivement les

vitesses rapides de convergence d/n. Nous discutons de l' optimalité de ces différentes vitesses à abaisser les limites en soulignant des références quand elles sont disponibles. En particulier, ces résultats apportent une généralisation des résultats de Dalalyan and Tsybakov (2008) sur l'estimation en régression sparse à certains auto-régression.

Abstract

This thesis aims at a systematic introduction to a weak dependence condition, provided by Doukhan and Louhichi (1999), which is more general than the classical frameworks of mixing or associated sequences. The notion is broad enough to include some standard models such as stable Markov models, Bilinear models, and more generally, Bernoulli shifts. In some cases no mixing properties can be expected without additional regularity assumption on the distribution of the innovations for which a weak dependence condition can be easily derived. We investigate the relationship between weak dependence and mixing for discrete valued processes. We show that weak dependence implies mixing conditions under natural assumptions. The results specialize to the case of Markov processes. Several examples of integer valued processes are discussed and their weak dependence properties are investigated by means of a contraction principle.

In the second part, we establish rates of convergences in statistical learning for time series forecasting. Using the PAC-Bayesian approach, slow rates of convergence $\sqrt{d/n}$ for the Gibbs estimator under the absolute loss were given in a previous work Alquier and Wintenberger (2012), where n is the sample size and d the dimension of the set of predictors. Under the same weak dependence conditions, we extend this result to any convex Lipschitz loss function. We also identify a condition on the parameter space that ensures similar rates for the classical penalized ERM procedure. We apply this method for quantile forecasting of the French GDP. Under additional conditions on the loss functions (satisfied by the quadratic loss function) and for uniformly mixing processes, we prove that the Gibbs estimator actually achieves fast rates of convergence d/n. We discuss the optimality of these different rates pointing out references to lower bounds when they are available. In particular, these results bring a generalization of the results of Dalalyan and Tsybakov (2008) on sparse regression estimation to some autoregression.

Contents

Chapitre 1

Introduction Générale et Résultats Principaux

Cette thèse porte sur l'inférence de la dépendance faible et la prévision des séries temporelles par l'approche PAC-bayésienne. Elle se compose de deux parties.

Le but de la première partie est d'étudier un système de dépendance faible. Nous donnons des grandes classes de modèles de séries temporelles qui satisfaient cette notion. Nous étudions la relation entre dépendance faible et mélangeant pour les processus de valeurs discrètes.

Cette première partie correspond au Chapitre 1 2 3. Chapitre 2 est constitué de l'article suivant :

1. On weak dependence conditions : The case of discrete valued processes, en collaboration avec Paul Doukhan et Konstantinos Fokianos, *Statistics and Probability Letters*, **82** (2012), 1941-1948.

La deuxième partie correspond aux Chapitres 4, dans lesquels on étudie les problèmes de prévision des séries temporelles. Cette seconde partie est constituée essentiellement de 2 articles :

2. Prediction of Quantiles by Statistical Learning and Application to GDP Forecasting, en collaboration avec Pierre Alquier, *in the proceedings of DS'12 (conference on Discovery Science), J.-G. Ganascia, P. Lenca and J.-M.*

Petit Eds., Springer - Lecture Notes in Artificial Intelligence, **7569** (2012), 22-36 ;

3. Prediction of Time Series by Statistical Learning : General Losses and Fast Rates, en collaboration avec Pierre Alquier et Olivier Wintenberge, *Dependence Modeling*, **Volume 1** (2013), 65-93.

Au cours des cinquante dernières années, diverses conditions de dépendance sont apparus dans la littérature, à la suite de la notion de mélange introduit par Rosenblatt (voir Rosenblatt (1985) pour plus d'information). Les notions de mélange ont été appliqués à de nombreux problèmes de type dépendant ; en particulier dans le contexte de séries temporelles et de leurs applications financières qui ont été appliqués à prouver des théorèmes limites qui permettent de valider l'inférence asymptotique ; voir Doukhan (1994), Rio (2000) et Bradley (2007) pour d'autres exemples. Cependant, pour certains modèles apparus fréquemment dans les applications, les conditions de mélange forts ne sont pas satisfaits. Les principaux exemples de ces modèles sont le célèbre AR (1) non-mélangeant modèle deAndrews (1984) et le LARCH(1) modèle considéré par Doukhan et al. (2006). Ces types de problèmes ont motivés Doukhan and Louhichi (1999) à introduire des conditions de dépendance plus flexible pour accueillir le plus grandes classes de modèles de séries temporelles. La principale notion introduite est que la dépendance faible ; le sujet est étudié de façon approfondie dans la monographie récente Dedecker et al. (2007) qui inclut de nombreux exemples de processus faiblement dépendantes.

Doukhan and Louhichi (1999) ont introduit un concept de dépendance faible pour les séries temporelles qui généralise les notions de mélange et association. Les covariances des variables aléatoirs sont beaucoup plus facile à calculer que les coefficients de mélange. Par conséquent la dépendance faible définie dans la définition 2.1.1 est mesurée en termes de covariances des fonctions. Supposons que, pour les fonctions commodes h et k,

$$\text{Cov } (h(\text{`past'}), k(\text{`future'}))$$

converge vers 0 comme la distance entre le "passé" et le "futur" converge vers l'infini. La convergence n'est pas supposé tenir uniformément sur la dimension du "passé" ou "futur" impliqués. Cette définition rend explicite l'indépendance

asymptotique entre le "passé" et le "futur" ; cela signifie que le "passé" est progressivement oublié.

Considérons $(X_t)_{t\in\mathbb{Z}}$ un processus à valeurs dans un espace $\mathbb{E}^u$ et $\|\cdot\|$ la norme correspondante. Nous définissons le module d'une fonction Lipschitzienne $h : \mathbb{E}^u \to \mathbb{R}$

$$\operatorname{Lip} h = \sup_{(y_1,\ldots,y_u)\neq(x_1,\ldots,x_u)\in E^u} \frac{|h(y_1,\ldots,y_u) - h(x_1,\ldots,x_u)|}{\|y_1 - x_1\| + \ldots + \|y_u - x_u\|}.$$

Définition 2.1.1. Soit $(X_t)_{t\in\mathbb{Z}}$ un processus à valeurs dans E. Soit $\Gamma(u, v, r)$ est l'ensemble des (s, t) en $\mathbb{Z}^u \times \mathbb{Z}^v$ tels que $s_1 \leq \ldots \leq s_u \leq s_u + r \leq t_1 \ldots \leq t_v$. Pour certains classes de functions $E^u, E^v \to \mathbb{R}, \mathcal{F}_u$, $\mathcal{G}_v$ le coefficient de dépendance est définit par

$$\epsilon(r) = \sup_{u,v} \sup_{(s,t)\in\Gamma(u,v,r)} \sup_{f\in\mathcal{F}_u, g\in\mathcal{G}_v} \frac{|\mathrm{Cov}\ (f(X_{s_1},\ldots,X_{s_u}), g(X_{t_1},\ldots,X_{t_v}))\,|}{\psi(f, g, u, v)}.$$

X_t est appelé processus (ϵ, ψ)-faiblement dependent si la séquence $\epsilon(r) \to_{r\to\infty} 0$.

Exemples d'intérêt concernent la fonction $\psi_1(f, g, u, v) = v\mathrm{Lip}\, g$ (par exemple dans les processus linéaires causal), $\psi_2(f, g, u, v) = u\mathrm{Lip}\, f + v\mathrm{Lip}\, g$, (par exemple dans les processus linéaire non causal), $\psi_3(f, g, u, v) = uv\mathrm{Lip}\, f \cdot \mathrm{Lip}\, g$ (par exemple dans les processus associés), et $\psi_4(f, g, u, v) = u\mathrm{Lip}\, f + v\mathrm{Lip}\, g + v\mathrm{Lip}\, f \cdot \mathrm{Lip}\, g$. Cette définition est héréditaire.

Il y a deux raisons pour lesquelles nous préférons utiliser la dépendance faible au lieu de mélange. Tout d'abord , les conditions de mélange se réfèrent plutôt à σ-algèbre qu'à des variables aléatoires. Ils sont donc plus adaptées à travailler dans des domaines de la finance ou d'histoire, où la σ-algèbre engendrée par le passé a une importance considérable. Deuxiémement, une difficulté de mélange est la vérification car il est généralement difficile (voir par exemple Doukhan (1994)), cependant, la dépendance faible a explicité un exemple simple d'un processus autoregressive avec des innovations de Bernoulli(Andrews (1984)) et a prouvé que ce modèle n'est pas fortement mélangeant , Doukhan and Louhichi (1999) ont montré que ce processus est faiblement dépendante. Cette notion de dépendance faible est suffisamment large pour inclure des nombreux exemples intéressants tels que les modèles de Markov stationnaires, modèles bilinéaires, et plus généralement, les schámas de Bernoulli. Plus précisément, dans

des conditions faibles, tous les processus causals ou non causals sont faiblement dépendants : par exemple les processus Gaussian, associés, linéaire, ARCH (∞), bilinéaires, Volterra, et les processus de mémoire infinie...

Nous discutons et étudions la relation entre le mélange et la dépendance faible pour les modèles temporelles à valeur entière. Au cours des dernières années, il y a une littérature émergente sur le thème de la modélisation et l'inférence pour les séries temporelles disctrètes, voir Kedem and Fokianos (2002),Doukhan et al. (2006), Drost et al. (2008) Fokianos et al. (2009), Fokianos and Tjøstheim (2011), Franke (2010) et Neumann (2011) pour modèles autorégressifs à valeur entière et modèles autoregressive généralisées. Nous allons nous concentrer sur ces modèles, mais nous signalons que d'autres familles de processus pourraient être considérés ; voir Coupier et al. (2006) pour le cas d'un processus général avec deux valeurs.

Notre objectif est de relier le mélange et la dépendance faible pour des modèles de séries temporelles à valeur entière. En utilisant la définition de η, la dépendance de la séquence $X_t, t \in \mathbb{Z}$ entre le passé et ses futurs r-uplets peut être évaluée suivant :

$$\left|\mathrm{Cov}\left(f(X_{i_1},\ldots,X_{i_u}), g(X_{j_1},\ldots,X_{j_v})\right)\right| \leq (u\mathrm{Lip}\, f + v\mathrm{Lip}\, g)\eta(r).$$

En raison du fait que les σ-algèbres générés par des ensembles discrets sont assez petites, nous montrons que les coefficients obtenus du monde de mélange coïncident souvent à ceux introduits sous dépendance faible. Par exemple, nous relions le coefficient de dépendance faible η pour les coefficients de mélange α.

Définition 3.3.1. Pour tout $d \geq 1$ on note $\|\cdot\|$ la norme uniforme, c'est à dire $\|(u_1,\ldots,u_d)\| = \max_{1\leq j\leq d}|u_j|$ sur $\mathbb{R}^d$. un ensemble $\mathbb{G}$ sera appelé discret si $\mathbb{G} \subset \mathbb{R}^d$ pour certains $d \geq 1$ et ses éléments satisfont

$$D = \inf_{x\neq x', x, x'\in\mathbb{G}} \|x - x'\| > 0$$

Proposition 3.3.1. Si $\{X_t, t \in \mathbb{Z}\}$ est un processus à valeur entière η-faiblement dépendant, alors

$$\alpha_{u,v}(r) \leq \frac{2}{D}(u+v)\eta(r)$$

Résultats analogues lorsque $X_t, t \in \mathbb{Z}$ est un processus à valeur entière τ-faiblement dépendante.

Le cas des processus de Markov est d'un intérêt particulier pour notre étude. Nous montrons que les coefficients de dépendance mènent une attention particulière aux chaînes de Markov. Plusieurs exemples de modèles autorègressifs entiers sont discutés en détail. En particulier, nous allons démontrer des conditions où les modèles existants doivent satisfaire de sorte qu'ils sont faiblement dépendants.

Le problème de prévision des séries temporelles est un problème fondamental dans les études de statistique. L'approche paramétrique contient une large famille de modèles associés à des méthodes d'estimation efficace et de prévision, voir par exemple Hamilton (1994); Brockwell and Davis (2009). Les modèles paramétriques classiques contiennent les processus linéaires tels que l'ARMA, et plus récemment, les processus non linéaires tels que les modèles volatilité stochastique et ARCH a reçu beaucoup d'attention dans les applications financières, voir, e.g., le papier séminal par le prix Nobel Engle (1991), et Francq and Zakoian (2010) pour une introduction plus récente. Cependant, dans la pratique, les hypothèses paramétriques tiennent rarement. Cela peut conduire à des prédictions très biaisée, et sous-évaluer les risques, voir Taleb (2007).

Au cours des dernières années, plusieurs approches universelles sont apparus dans divers domaines tels que la statistique non paramétrique, l'apprentissage automatique, l'informatique et la théorie de jeux. Ces approches partagent certaines caractéristiques communes : l'objectif est de construire une procédure qui prévoit la série ainsi que le meilleur prédicteur dans un ensemble donné de variables prédictives initiales, sans aucune hypothèse paramétrique de la distribution de l'observation. Cependant, l'ensemble de prédicteurs peut être inspiré par différents modèles statistiques paramètriques ou non paramétriques. Nous pouvons distinguer deux des classes de ces approches, avec quantification différent de l'objectif, et des terminologies différentes :

- dans l'approche "prédiction de séquences individuelles", les facteurs prédictifs sont généralement appelés des "experts". L'objectif est la prd́iction en ligne : à chaque date t, une prédiction de la réalisation de l'avenir x_{t+1} est basée sur l'observations précédente x_1,...,x_t, l'objectif est de minimiser la

perte de prédiction cumulative. Voir par exemple Cesa-Bianchi and Lugosi (2006); Stoltz (2009) pour une introduction.

- dans l'approche de l'apprentissage statistique, les prédicteurs proposés sont parfois appelés "modèles" ou des "concepts". Le cadre de batch est plus classique dans la statistique. Une procédure de la prédiction est construite sur un échantillon complet X_1,..., X_n. La performance de la procédure est comparée à la moyenne avec le meilleur prédicteur, appelé "l'oracle". L'environnement n'est pas déterministe et certaines hypothèses comme mélange ou dépendance faible sont nécessaires : voir Meir (2000); Modha and Masry (1998); Alquier and Wintenberger (2012). Notez que les résultats de l'approche "prédiction de séquences individuelles" peuvent généralement être étendus à ce cadre, voir par exemple Gerchinovitz (2011) pour le cas iid, et Agarwal and Duchi (2011); Agarwal et al. (2012) pour le mélange de la série chronologique.

Dans les deux cas, on est généralement capables de prévoir une série temporelle borné ainsi que le meilleur expert, jusqu'à un petit residu Δ_n. Ce type de résultats est appelé dans la théorie de statistique une inégalité d'oracle. En général, en négligeant la taille de l'ensemble des prédicteurs θ, le residu est de l'ordre $1/\sqrt{n}$ dans les deux approches : voir par exemple Cesa-Bianchi and Lugosi (2006) pour l'approche "séquences individuelles", pour l'approche "statistique de l'apprentissage" la vitesse $1/\sqrt{n}$ est atteinte dans Alquier and Wintenberger (2012) avec la fonction de perte absolue et sous une hypothése de dépendance faible . Différentes procédures sont utilisées pour atteindre ces vitesses. Citons la minimisation du risque empiriqueVapnik (1999) et les procédures d'agrégation avec des poids exponentiels, souvent référé comme l'EWA Dalalyan and Tsybakov (2008); Gerchinovitz (2011) ou l'estimateur Gibbs Catoni (2004, 2007), lié à l'approche en ligne de l'algorithme pondération majoritaire Littlestone and Warmuth (1994), see also Vovk (1990).

Dans cette thèse, nous nous concentrons sur la prévision des séries temporelles en utilisant l'approche de l'apprentissage statistique. Soit X_1,..., X_n représentent des observations effectuées à temps $t \in \{1, \ldots, n\}$ de la série temporelle $X = (X_t)_{t\in\mathbb{Z}}$ définie sur $(\Omega, \mathcal{A}, \mathbb{P})$. Nous supposons que cette série temporelle prend des valeurs dans $\mathbb{R}^p$ équipés de la norme euclidienne $\|\cdot\|$. Comme mentionné

ci-dessus, dans l'approche de la théorie de l'apprentissage, fixe un entier k, nous supposons que l'on nous donne un ensemble de prédicteurs

$$\{f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta\}$$

où Θ est un sous-ensemble de l'espace linéaire pour des raisons de simplicité. Toutefois, le θ ici représente l'union de tous les paramètres de tous les modèles que nous envisageons. Nous allons utiliser une approche du type sélection de modèle :

$$\Theta = \cup_{j=1}^{M} \Theta_j.$$

Θ sera une union finie (ou plus généralement dénombrable) de sous-espaces. L'importance de l'introduction une telle structure a été mise en avant par Vapnik (1999), c'est un moyen d'éviter de faire des hypothèses fortes sur la distribution des observations.

Dans l'approche PAC-bayésienne, nous menons des prévisions de séries temporelle dans un contexte où les inégalités du type Hoeffding ou Bernstein peuvent être appliquées, puis à se débarrasser des échantillons d'observation par une intégration. Afin de mesurer la complexité de l'espace de paramètres θ, nous considérons un σ-algèbre $\mathcal{T}$ sur θ, soit $\mathcal{M}_+^1(\theta)$ représentent l'ensemble de toutes les mesures de probabilité sur $(\Theta, \mathcal{T})$, nous définissons une distribution de probabilité $\pi \in \mathcal{M}_+^1(\theta)$. Remarquons que π est aussi appelée la distribution a priori dans le point de vue PAC-bayésien, mais ne dispose pas d'interprétation bayésienne. Plus précisément, π ne tient pas compte de toute croyance préalable sur la localisation de la "vraie" valeur du paramètre ni modélisation stochastique de $\theta \in \Theta$, π joue juste le rôle de définir une structure en Θ liés à la mesure de la complexité de θ.

Les bornes PAC-bayésiens ont été introduits dans Shawe-Taylor and Williamson (1997); McAllester (1999) dans le contexte de $0-1$ classification, Il peut traiter des problèmes très généraux et donne des résultats sur le choix du modèle et de l'agrégation, voir Catoni (2004, 2007); Alquier (2008); Audibert (2010); Audibert and Catoni (2011) pour les traveaux les plus rcents. Le nom est en raison du fait que, dans sa première forme son objectif était de combiner les principaux avantages du point de vue de théorie de l'apprentissage et des statistiques bayésiens. Dans l'apprentissage statistique, les bornes sur le risque $R(\hat{\theta})$

d'un estimateur $\hat{\theta}$ souvent dépend du risque empirique de $\hat{\theta}$, $r_n(\hat{\theta})$, et sur une mesure de la complexité de la sous-modèle de Θ utilisé pour construire $\hat{\theta}$.

La technique utilisée dans cette thèse est inspirée par celle mise au point récemment par Catoni (2004, 2007). Il utilise une structure d'une distribution "préalable" de probabilité sur l'espace de paramètre Θ : $\pi \in \mathcal{M}_+^1(\theta)$ pour remplacer la structure de sous-modèles de Θ. Au lieu de borner le minimum du risque empirique par rapport au paramètre $\theta \in \Theta$, nous étudions les déviation des quantiles de $r_n(\theta)$ par rapport à une mesure de probabilité a priori $\pi \in \mathcal{M}_+^1(\theta)$ définie sur l'espace des paramètres.

L'idée de l'approche PAC-bayésienne est que le risque de l'estimateur de Gibbs sera proche de $\inf_\theta R(\theta)$ jusque'à un petit résidu qui est remplacé par une mesure de la distance entre ρ et π. Pour des raisons de simplicité, nous posons $\overline{\theta} \in \Theta$ telle que

$$R(\overline{\theta}) = \inf_{\theta \in \Theta} R(\theta)$$

(Si un tel minimiseur n'existe pas, nous pouvons le remplacer par un minimiseur approximative $R(\overline{\theta}_\alpha) \leq \inf_\theta R(\theta) + \alpha$).

Dans le point de vue de PAC-bayésienne, on est généralement capable de prévoir une série temporelle aussi bien que le meilleur modèle ou expert, jusque'à un terme d'erreur qui diminue avec le nombre d'observations n. Ce type de résultats est appelé les inégalités oracle dans la théorie statistique. Autrement dit, on construit un prédicteur $\hat{\theta}$ sur la base des observations de telle sorte que

$$R(\hat{\theta}) \leq \inf_{\theta \in \Theta} R(\theta) + \Delta(n, \Theta)$$

où $R(\theta)$ est une mesure du risque de prédiction du prédicteur $\theta \in \Theta$. En général, le terme de residu est de l' order $\Delta(n, \Theta) \sim \sqrt{c(\Theta)/n}$, où $c(\Theta)$ mesure la complexité de Θ. Ici, cela se fait avec la divergence de Kullback :

$$\mathcal{K}(\rho, \pi) = \rho \left[\log \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \right) \right]$$

si ρ est absolument continue par rapport à π, sinon $\mathcal{K}(\rho, \pi) = \infty$.

Nous vous présentons un premier exemple de ce type de résultats prśentés dans cette thèse.

Théorème 5.5.1. Supposons que **LowRates(κ)** est satisfait pour certains $\kappa > 0$. Alors, pour tout λ, $\varepsilon > 0$, avec la probabilité au moins $1 - \varepsilon$ on obtient

$$R\left(\hat{\theta}_\lambda\right) \qquad \leq \qquad \inf_{\rho\in\mathcal{M}_+^1(\Theta)}\left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\mathcal{K}(\rho,\pi)+2\log\left(2/\varepsilon\right)}{\lambda}\right].$$

Le choix du paramètre λ est un problème difficile dans le cadre de la dépendance, c'est discuté en détail dans cette thèse. De plus, sous des hypothèses supplémentaires sur le modèle, nous pouvons montrer que la procédure classique de Minimisation du Risque empirique(ERM) peut être utilisé à la place de l'estimateur Gibbs. Au contraire de l'estimateur de Gibbs, il n'y a pas de paramètre de réglage pour l'ERM, donc c'est une situation très favorable.

Cependant, il est connu que, en théorie d'apprentissage si l'on veut avoir en fait à réaliser des estimateurs atteignent effectivement une vitesse rapide de convergence d/n, les théorèmes comme 5.5.1 ne sont pas suffisantes. Dans des conditions supplémentaires sur les fonctions de perte (satisfaites par la fonction de perte quadratique) et pour les processus de mélange uniforme, nous montrons dans cette thèse que la vitesse $1/n$ peut être atteinte.

Théorème 5.7.1. Supposons que :

1. **Margin**$(\mathcal{K})$ et **LipLoss**(K) sont satisfaits pour certains K, $\mathcal{K}>0$;
2. **PhiMix**$(\mathcal{B},\mathcal{C})$ est satisfait pour certains $\mathcal{C}>0$;
3. **Lip**(L) est satisfait pour certains $L>0$;
4. pour tout $j\in\{1,...,M\}$, il existe $d_j=d(\Theta_j,\pi)$ et $R_j=R(\Theta_j,\pi_j)$ satisfaisant la relation

$$\forall\delta>0,\quad \log\frac{1}{\int_{\theta\in\Theta_j}\mathbb{1}\{R(\theta)-R(\overline{\theta}_j)<\delta\}\pi_j(\mathrm{d}\theta)}\leq d_j\log\left(\frac{D_j}{\delta}\right).$$

Alors pour

$$\lambda=\frac{n-k}{4kKL\mathcal{B}\mathcal{C}}\wedge\frac{n-k}{16k\mathcal{C}}$$

pour tout $\varepsilon>0$, l'inequalité oracle (5.3) est avec

$$\Delta(n,\lambda,\pi,\varepsilon)$$
$$=4\inf_j\left\{R(\overline{\theta}_j)-R(\overline{\theta})+4k\mathcal{C}\left(4\vee KL\mathcal{B}\right)\frac{d_j\log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{n-k}\right\}.$$

Notons que Agarwal and Duchi (2011) prouve la vitesse rapide pour les algorithmes en ligne qui sont également de calcul efficace, voir aussi Agarwal et al. (2012). La vitesse rapide $1/n$ est atteinte lorsque les coefficients (ϕ_r) sont géométriquement diminués. Dans d'autres cas, la vitesse est plus lente. Nous ne souffrons pas d'une telle restriction ici. Il faut noter que les algorithmes efficaces de Monte Carlo sont disponibles pour calculer ces estimateurs de poids exponentiels , voir par exemple Alquier and Lounici (2011); Dalalyan and Tsybakov (2008).

General Introduction

Over the last fifty years or so, various dependence conditions have emerged in literature, as a result of the notion of mixing introduced by Rosenblatt (see Rosenblatt (1985) for more). Mixing notions have been applied to numerous dependence type problems; especially in the context of time series and their financial applications they were applied on proving limit theorems which enable valid asymptotic inference; see Doukhan (1994), Rio (2000) and Bradley (2007) for further examples. However, for some models encountered frequently in applications, strong mixing conditions are not satisfied. Prominent examples of such models are the celebrated AR(1) non-mixing model of Andrews (1984) and the LARCH(1) model considered by Doukhan et al. (2006). These types of problems motivated Doukhan and Louhichi (1999) to introduce more flexible dependence conditions to accommodate larger classes of time series models. The main notion introduced is that of weak dependence; the topic is studied extensively in the recent monograph by Dedecker et al. (2007) which includes numerous examples of weakly dependent processes.

Doukhan and Louhichi (1999) have introduced a concept of weak dependence for time series which generalizes the notions of mixing and association. Covariances of r.v.s are much easier to compute than mixing coefficients. Therefore weak dependence as defined in Definition 2.1.1 is measured in terms of covariances of functions. Assume that, for convenient functions h and k,

$$\text{Cov } (h(\text{'past'}), k(\text{'future'}))$$

converge to 0 as the distance between the 'past' and the 'future' converges to infinity. The convergence is not assumed to hold uniformly on the dimension of the 'past' or 'future' involved. This definition makes explicit the asymptotic independence between 'past' and 'future'; this means that the 'past' is progressively forgotten. Consider $(X_t)_{t\in\mathbb{Z}}$ a process with values in some space $\mathbb{E}^u$ and $\|\cdot\|$ the corresponding norm. We define the Lipschitz modulus of a function $h : \mathbb{E}^u \to \mathbb{R}$

$$\text{Lip}\, h = \sup_{(y_1,\ldots,y_u)\neq(x_1,\ldots,x_u)\in E^u} \frac{|h(y_1,\ldots,y_u) - h(x_1,\ldots,x_u)|}{\|y_1 - x_1\| + \ldots + \|y_u - x_u\|}.$$

Definition 2.1.1. Let $(X_t)_{t\in\mathbb{Z}}$ be a process with values in E. Let $\Gamma(u, v, r)$ be the set of (s, t) in $\mathbb{Z}^u \times \mathbb{Z}^v$ such that $s_1 \leq \ldots \leq s_u \leq s_u + r \leq t_1 \ldots \leq t_v$. For some

classes of functions $E^u, E^v \to \mathbb{R}, \mathcal{F}_u$, $\mathcal{G}_v$ the dependence coefficient is defined by

$$\epsilon(r) = \sup_{u,v} \sup_{(s,t)\in\Gamma(u,v,r)} \sup_{f\in\mathcal{F}_u, g\in\mathcal{G}_v} \frac{|\text{Cov } (f(X_{s_1},\ldots,X_{s_u}), g(X_{t_1},\ldots,X_{t_v}))\,|}{\psi(f,g,u,v)}.$$

X_t is called (ϵ,ψ)-weakly dependent process if the sequence $\epsilon(r) \to_{r\to\infty} 0$.

Examples of interest involve the function $\psi_1(f,g,u,v) = v\text{Lip}\, g$ (e.g. in causal linear processes), $\psi_2(f,g,u,v) = u\text{Lip}\, f + v\text{Lip}\, g$, (e.g. in non causal linear processes), $\psi_3(f,g,u,v) = uv\text{Lip}\, f \cdot \text{Lip}\, g$ (e.g. in associated processes), and $\psi_4(f,g,u,v) = u\text{Lip}\, f + v\text{Lip}\, g + v\text{Lip}\, f \cdot \text{Lip}\, g$. This definition is hereditary through images by convenient functions.

There are two reasons we prefer using weak dependence instead of mixing. Firstly, mixing conditions refer rather to σ-algebra than to random variables. They are consequently more adapted to work in areas like Financ, that is the σ-algebra generated by the past is of a considerable importance. Secondly, A difficulty of mixing is that checking for it is usually hard.(see e.g Doukhan (1994)) however, weak dependence is a very general property including certain non-mixing processes: e.g. Andrews (1984) explicated the simple example of an autoregressive process with Bernoulli innovations and proved that such a model is not strong mixing, while Doukhan and Louhichi (1999) proved that such a process is weakly dependent.

This weak dependence notion is broad enough to include many interesting examples such as stationary Markov models, bilinear models, and more generally, Bernoulli shifts. More precisely, under weak conditions, all the usual causal or non causal time series are weakly dependent processes: this is the case for instance of Gaussian, associated, linear, ARCH(∞), bilinear, Volterra, infinite memory processes, . . .

We discuss and investigate the relationship between mixing and weak dependence for integer valued time series models. In recent years, there is an emerging literature on the topic of modeling and inference for count time series, see Kedem and Fokianos (2002), Doukhan et al. (2006), Drost et al. (2008) Fokianos et al. (2009), Fokianos and Tjøstheim (2011), Franke (2010) and Neumann (2011) for integer autoregressive models and for generalized autoregressive models, among other references. We will focus on such models but we point out that other families might be considered as well; see Coupier et al. (2006) for the case of a general process with two values.

Our objective is to relate mixing and weak dependence conditions for such integer valued count time series models. Using the definition of η, the dependence between the past of the sequence $X_t, t \in \mathbb{Z}$ and its future r-tuples may be assessed as follows.

$$\left|\mathrm{Cov}\left(f(X_{i_1}, \ldots, X_{i_u}), g(X_{j_1}, \ldots, X_{j_v})\right)\right| \leq (u\mathrm{Lip}\, f + v\mathrm{Lip}\, g)\eta(r).$$

Due to the fact that the σ-algebras generated by discrete sets are quite small, we prove that the coefficients obtained from the mixing world often coincide to those introduced under weak dependence. For example, we link the weak dependence coefficient η to the strong mixing coefficients α.

Definition 3.3.1. For each $d \geq 1$ we denote by $\|\cdot\|$ the uniform norm, i.e. $\|(u_1, \ldots, u_d)\| = \max_{1\leq j\leq d} |u_j|$ on $\mathbb{R}^d$. A set $\mathbb{G}$ will be called discrete if $\mathbb{G} \subset \mathbb{R}^d$ for some $d \geq 1$ and its elements satisfy

$$D = \inf_{x\neq x', x, x'\in\mathbb{G}} \|x - x'\| > 0$$

Proposition 3.3.1. If $\{X_t, t \in \mathbb{Z}\}$ is an η-weakly dependent integer valued process, then

$$\alpha_{u,v}(r) \leq \frac{2}{D}(u + v)\eta(r)$$

Similar results when $\{X_t, t \in \mathbb{Z}\}$ is a τ-weakly dependent integer valued process.

The case of Markov processes is of a particular interest in our investigation. We show that the various coefficients of dependence lead special attention to Markov chains. Several examples of integer autoregressive models are discussed in detail. In particular, we will prove conditions which existing models should satisfy so that they are weakly dependent.

The problem of time series forecasting is a fundamental problem in statistics. The parametric approach contains a wide range of models associated with efficient estimation and prediction methods, see e.g. Hamilton (1994); Brockwell and Davis (2009). Classical parametric models include linear processes such

as ARMA, and more recently, non-linear processes such as stochastic volatility models and ARCH received a lot of attention in financial applications - see e.g. the seminal paper by Nobel prize winner Engle (1991), and Francq and Zakoian (2010) for a more recent introduction. However, in practice, parametric assumptions rarely holds. This can lead to highly biased prediction, and to underevaluate the risks, see among others the polemical but highly informative discussion in Taleb (2007).

In the last few years, several universal approaches emerged from various fields such that non-parametric statistics, machine learning, computer science and game theory. These approaches share some common features: the aim is to build a procedure that predicts the series as well as the best predictor in a given set of initial predictors, without any parametric assumption on the distribution of the observation. However, the set of predictors can be inspired by different parametric or non-parametric statistical models. We can distinguish two classes in these approaches, with different quantification of the objective, and different terminologies:

- in the "prediction of individual sequences" approach, predictors are usually called "experts". The objective is online prediction: at each date t, a prediction of the future realization x_{t+1} is based on the previous observations x_1, ..., x_t, the objective being to minimize the cumulative prediction loss. See for example Cesa-Bianchi and Lugosi (2006); Stoltz (2009) for an introduction.

- in the statistical learning approach, the given predictors are sometimes referred to as "models" or "concepts". The batch setting is more classical in statistics. A prediction procedure is build on a complete sample X_1, ..., X_n. The performance of the procedure is compared on average with the best predictor, called the 'oracle". The environment is not deterministic and some hypotheses like mixing or weak dependence is required: see Meir (2000); Modha and Masry (1998); Alquier and Wintenberger (2012). Note that results from the "individual sequences" approach can usually be extended to this setting, see e.g. Gerchinovitz (2011) for the iid case, and Agarwal and Duchi (2011); Agarwal et al. (2012) for mixing time series.

In both settings, one is usually able to predict a bounded time series as well

as the best expert, up to a small remainder Δ_n. This type of results is referred in statistical theory as an oracle inequality. In general, neglecting the size of the set of predictors Θ, the remainder is of the order $1/\sqrt{n}$ in both approaches: see e.g. Cesa-Bianchi and Lugosi (2006) for the "individual sequences" approach; for the "statistical learning approach" the rate $1/\sqrt{n}$ is reached in Alquier and Wintenberger (2012) with the absolute loss function and under a weak dependance assumption. Different procedures are used to reach these rates. Let us mention the empirical risk minimization Vapnik (1999) and aggregation procedures with exponential weights, usual referred as EWA Dalalyan and Tsybakov (2008); Gerchinovitz (2011) or Gibbs estimator Catoni (2004, 2007) in the batch approach, linked to the weighted majority algorithm of the online approach Littlestone and Warmuth (1994), see also Vovk (1990).

In this thesis, we focus on the time series forecasting using the statistical learning approach. Let $X_1, \dots, X_n$ denote the observations at time $t \in \{1, \dots, n\}$ of a time series $X = (X_t)_{t\in\mathbb{Z}}$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$. We assume that this time series takes values in $\mathbb{R}^p$ equipped with the Euclidean norm $\|\cdot\|$. As mentioned above, in the learning theory approach, fixed an integer k, we assume that we are given a set of predictors

$$\{f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta\}$$

where Θ is subset of a linear space for the sake of simplicity. However the Θ here represent the union of all the parameters of all the models we envision. We will use a model-selection type approach:

$$\Theta = \cup_{j=1}^{M} \Theta_j.$$

Θ will be a finite (or more generally countable) union of subspaces. The importance of introducing such a structure has been put forward by Vapnik (1999), as a way to avoid making strong hypotheses on the distribution of the observations.

In the PAC-Bayesian approach, we lead time series forecasting to a context where Hoeffding or Bernstein type inequalities can be applied, and then to get rid of the observation samples by an integration with respect it. In order to measure the complexity of the parameter space Θ, we consider a σ-algebra $\mathcal{T}$ on Θ, let $\mathcal{M}_+^1(\Theta)$ denote the set of all probability measure on $(\Theta, \mathcal{T})$, we define a probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$. Remark that π is also called the prior distribution in the PAC- Bayesian point of view but does not have any Bayesian

interpretation. More precisely, π does not reflect any prior belief on the localization of the "true" value of the parameter nor a stochastic modelization of $\theta \in \Theta$, π just plays the role on defining a structure over Θ involved in measuring the complexity of Θ.

PAC-Bayesian bounds were introduced in Shawe-Taylor and Williamson (1997); McAllester (1999) in the context of $0-1$ classification, It can deal with very general problems and gives results about model selection and aggregation, see Catoni (2004, 2007); Alquier (2008); Audibert (2010); Audibert and Catoni (2011) for more recent advances. Its name is due to the fact that in its first form its objective was to combine the major advantages of the learning theory point of view and of the Bayesian statistics. In statistical learning, the bounds on the risk $R(\hat{\theta})$ of an estimator $\hat{\theta}$ often depends on the empirical risk of $\hat{\theta}$, $r_n(\hat{\theta})$, and on a measure of the complexity of the submodel of Θ used to build $\hat{\theta}$.

The technique used in this thesis is inspired by the one developed more recently by Catoni (2004, 2007). He uses as a structure a "prior" probability distribution over the parameter space Θ: $\pi \in \mathcal{M}_+^1(\Theta)$ to replace the structure of submodels of Θ. Instead of bounding the minimum of the empirical risk with respect to the parameter $\theta \in \Theta$, we study the deviations of the quantiles of $r_n(\theta)$ with respect to some prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ defined on the parameter space. The idea of PAC-Bayesian approach is that the risk of the Gibbs estimator will be close to $\inf_\theta R(\theta)$ up to a small remainder which is replaced by a measure of the distance between ρ and the π. For the sake of simplicity, let $\overline{\theta} \in \Theta$ be such that

$$R(\overline{\theta}) = \inf_{\theta \in \Theta} R(\theta)$$

(if such a minimizer do not exist, we can just replace it by an approximate minimizer $R(\overline{\theta}_\alpha) \leq \inf_\theta R(\theta) + \alpha$).

In the PAC-Bayesian point of view, one is usually able to predict a time series as well as the best model or expert, up to an error term that decreases with the number of observations n. This type of results is referred to as oracle inequalities in statistical theory. In other words, one builds on the basis of the observations a predictor $\hat{\theta}$ such that

$$R(\hat{\theta}) \leq \inf_{\theta \in \Theta} R(\theta) + \Delta(n, \Theta)$$

where $R(\theta)$ is a measure of the prediction risk of the predictor $\theta \in \Theta$. In general, the remainder term is of the order $\Delta(n, \Theta) \sim \sqrt{c(\Theta)/n}$, where $c(\Theta)$ measures the complexity of Θ. Here, this is done with the Kullback divergence:

$$\mathcal{K}(\rho, \pi) = \rho \left[\log \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \right) \right]$$

if ρ is absolutely continuous with respect to π, otherwise $\mathcal{K}(\rho, \pi) = \infty$.

Let us introduce a first example of this kind of results presented in the second part of this thesis.

Theorem 5.5.1. Let us assume that **LowRates(κ)** is satisfied for some $\kappa > 0$. Then, for any λ, $\varepsilon > 0$, with probability at least $1 - \varepsilon$ we have

$$R\left(\hat{\theta}_\lambda\right) \quad \leq \quad \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\int R \mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + \frac{2\mathcal{K}(\rho, \pi) + 2\log\left(2/\varepsilon\right)}{\lambda} \right].$$

The choice of the parameter λ is a hard problem in the context of dependence, it is discussed in details in this thesis. Also, under additional assumptions on the model, we can prove that the classical Empirical Risk Minimization (ERM) procedure can be used instead of the Gibbs estimator. On the contrary to the Gibbs estimator, there is no tuning parameter for the ERM, so this is a very favorable situation.

However, it is a known fact that in learning theory that if one wants to have estimators actually achieve fast rates of convergence d/n, theorems like 5.5.1 are not sufficient. Under additional conditions on the loss functions(satisfied by the quadratic loss function) and for uniformly mixing processes, we prove in this thesis that the rate $1/n$ can be achieved.

Theorem 5.7.1. Assume that:

1. **Margin**($\mathcal{K}$) and **LipLoss**(K) are satisfied for some K, $\mathcal{K} > 0$;
2. **PhiMix**($\mathcal{B}, \mathcal{C}$) is satisfied for some $\mathcal{C} > 0$;
3. **Lip**(L) is satisfied for some $L > 0$;
4. for any $j \in \{1, ..., M\}$, there exist $d_j = d(\Theta_j, \pi)$ and $R_j = R(\Theta_j, \pi_j)$ satisfying the relation

$$\forall \delta > 0, \quad \log \frac{1}{\int_{\theta \in \Theta_j} \mathbb{1}\{R(\theta) - R(\bar{\theta}_j) < \delta\} \pi_j(\mathrm{d}\theta)} \leq d_j \log\left(\frac{D_j}{\delta}\right).$$

Then for

$$\lambda = \frac{n-k}{4kKL\mathcal{B}\mathcal{C}} \wedge \frac{n-k}{16k\mathcal{C}}$$

the oracle inequality (5.3) for any $\varepsilon > 0$ with

$$\Delta(n, \lambda, \pi, \varepsilon) = 4 \inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + 4k\mathcal{C}\,(4 \vee KL\mathcal{B})\, \frac{d_j \log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{n-k} \right\}.$$

Note that Agarwal and Duchi (2011) proves fast rates for online algorithms that are also computationally efficient, see also Agarwal et al. (2012). The fast rate $1/n$ is reached when the coefficients (ϕ_r) are geometrically decreasing. In other cases, the rate is slower. We do not suffer such a restriction here. It should be noted that efficient Monte Carlo algorithms are available to compute these exponential weights estimators, see for example Alquier and Lounici (2011); Dalalyan and Tsybakov (2008).

Part I

Weak Dependence, Models and Applications

Chapter 2

Weak Dependence Notions and Models

The aim of this part is to propose a mathematical introduction to the content of dependence. To do this, we recall weak dependence conditions from Dedecker et al. (2007) (Weak dependence, examples and applications. Lecture Notes in Statistics, Vol 190)'s monograph. Mixing sequences, functions of associated or Gaussian sequences, Bernoulli shifts as well as models with a Markovian representation are examples of the models considered. We investigate the relationship between mixing and weak dependence for integer valued time series models.

2.1 Introduction

We start here from some very basic facts concerning independence of random variables. We suppose that we are given P and F two random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Note $\sigma(P)$ the σ-algebra generated by P, and respectively $\sigma(F)$. So independence of both random variables writes as

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \ \forall A \in \sigma(P), \ \ \forall B \in \sigma(F).$$

This definition can be extended to

$$\mathrm{Cov}\left(f(P), g(F)\right) = 0$$

for all f,g with $\|f\|_\infty, \|g\|_\infty \leq 1$.

If now, we consider a time series $X = (X_t)_{t\in\mathbb{Z}}$, the variable P and F may be denoted 'Past' and 'Future':

$$P = (X_{i_1}, \ldots, X_{i_u}), \quad F = (X_{j_1}, \ldots, X_{j_v})$$

for $i_1 \leq i_2 \leq \ldots, \leq i_u < j_1 \leq j_2 \ldots \leq j_v$, $u, v \in \mathrm{N}^*$.

Since no phenomena are really independent from each others, a first question is asked here, how to weaken those relations.

A first answer to this problem was the mixing assumption introduced by Rosenblatt (1956). For a long time mixing conditions have been the dominant type of conditions for imposing a restriction on the dependence between time series data. They are considered to be useful since they are fulfilled for many classes of processes and since they allow us to derive tools similar to those in the independent case. However, mixing conditions can be very hard to verify for particular models or are even too strong to be true (see. e.g Doukhan (1994)) and such conditions refer rather to σ-algebras than to random variables.

Covariances of r.v.s are much easier to compute than mixing coefficient. Therefore Doukhan and Louhichi (1999) have introduced a concept of weak dependence to the case of time series which generalizes the notion of mixing and association. It is measured in terms of covariances of functions. For convenient functions h and k, we assume that

$$\text{Cov } (h(\text{'past'}), k(\text{'future'}))$$

is small when the distance between the 'past' and the 'future' is sufficiently large. This definition makes explicit the asymptotic independence of finite-dimensional distribution with separated index sets (see Definition 2.1.1); the convergence is not assumed to hold uniformly on the dimension of the distributions involved.

Wu (2005) introduced the physique dependence measures for stationary causal process. Based on the nonlinear system theory, he introduces dependence coefficients by measuring the degree of dependence of outputs on inputs in physical system. Asymptotic properties have been established under such dependence conditions.

2.1.1 Mixing

Mixing conditions are defined in terms of the σ-algebras generated by a random sequence.

$$\begin{aligned}
\alpha(\sigma(P),\sigma(F)) &= \sup_{P\in\sigma(P),F\in\sigma(F)} |\mathbb{P}(P)\mathbb{P}(F)-\mathbb{P}(P\cap F)| \\
\beta(\sigma(P),\sigma(F)) &= \|\mathbb{P}_{(P,F)}-\mathbb{P}_P\otimes\mathbb{P}_F\|_{TV} \\
\rho(\sigma(P),\sigma(F)) &= \sup_{p\in \mathrm{L}^2(\sigma(P)),f\in \mathrm{L}^2(\sigma(F))} |Corr(p,f)| \\
\phi(\sigma(P),\sigma(F)) &= \sup_{P\in\sigma(P),V\in\sigma(F)} |\frac{\mathbb{P}(P\cap F)}{\mathbb{P}(P)}-\mathbb{P}(F)|
\end{aligned}$$

The β-mixing coefficient, introduced by Wolkonski and Rozanov (1959, 1961),Kolmogorov and Rozanov (1978) introduced the maximal correlation coefficient ρ and defined the corresponding dependence condition. The coefficient ϕ is the uniform mixing coefficient by Ibragimov (1962).

Proposition 2.1.1. *The following relations hold:*

$$\phi-mixing \Rightarrow \left\{ \begin{array}{c} \rho-mixing \\ \beta-mixing \end{array} \right\} \Rightarrow \alpha-mixing$$

Proof is omitted and more details and examples for such conditions can be found in Doukhan (1994) and Rio (2000), there is no reverse implication holds in general.

As basic assumptions on the dependence structures, the mixing conditions have been widely used and various limit theorems have been obtained; It is impossible to give a complete list of references here. Representative results are Doukhan (1994), Rio (2000) and Bradley (2007). However, most of the asymptotic results developed in the literature are for strong mixing processes and processes with quite restrictive summability conditions on joint cumulants. Such conditions seem restrictive and they are not easily verifiable. For example, Andrews (1984) showed that, for a simple autoregressive process with innovations being independent and identically distributed (iid) Bernoulli random variables, the process is not strong mixing.

Example 2.1.1. *Andrews (1984)'s simple example is, however, not mixing*

$$X_t = \frac{1}{2}(X_{t-1} + \xi_t), \;\; \xi_t \sim b\left(\frac{1}{2}\right), \textit{iid.}$$

X_t has the uniform density over $(0,1)$. X_t is a causal process with the representation $X_t = \sum_{j=0}^{\infty} 2^{-j}\xi_{t-j}$ and the innovations $\xi_t, \xi_{t-1}, \ldots$ correspond to the dyadic expansion of X_t. ξ_{t-k} is the k-th digit in the binary expansion of the uniformly chosen number $X_t = 0.\xi_t\xi_{t-1}\ldots \in [0,1]$. This shows that X_0 is some deterministic function of X_t which derives that such models are not mixing. Thus the process X_t is not strong mixing and $\alpha_n \equiv 1/4$ for all t.

Example 2.1.2. *In Doukhan et al. (2009) paper, one extend andrew's idea and provide an $LARCH(1)$ not-mixing model:*

$$X_t = \xi_t(1 + aX_{t-1})$$

where $\mathbb{P}(\xi_0 = 1) = \mathbb{P}(\xi_0 = -1) = 1/2$.

This model has the stationary uniform distribution in $\mathbb{L}^m$ with $m \geq 1$,

$$X_t = \xi_t + \sum_{j \geq 1} a^j \xi_t \ldots \xi_{t-j}. \tag{2.1}$$

But it satisfies no mixing condition if $a \in (\frac{3-\sqrt{5}}{2}, \frac{1}{2}]$ (the past may entirely be recovered from the present).

The proof is as in Andrews (1984) that $\mathbb{P}(X_t \in A | X_{t-(n+1)} \in B) = 1, (\forall n)$, $\mathbb{P}(X_{t-(n+1)} \in B) \neq 0$ and $\mathbb{P}(X_t \in A) < 1$, for some well chosen subsets A, B of $\mathbb{R}$. Set $U_t = (X_t \in A)$ and $V_{t-n-1} = (X_{t-(n+1)} \in B)$ then $\mathbb{P}(U_t \cap V_{t-n-1}) = \mathbb{P}(V_{t-n-1})$ and we derive from stationarity that $\mathbb{P}(V_{t-n-1}) = \mathbb{P}(V_0) \neq 0$ and $\mathbb{P}(U_t) = \mathbb{P}(U_0) < 1$; thus $\alpha_n \geq \mathbb{P}(U_t \cap V_{t-n-1}) - \mathbb{P}(U_t)\mathbb{P}(V_{t-n-1}) \geq \mathbb{P}(V_0)(1 - \mathbb{P}(U_0)) > 0$.

We use the decomposition:

$$X_t = A_{t,n} + a^{n+1}\xi_t \ldots \xi_{t-n} X_{t-(n+1)}, \;\; A_{t,n} = \xi_t + a\xi_t\xi_{t+1} + \ldots + a^n \xi_t \ldots \xi_{t-n}.$$

1. The values of the random variable $A_{t,n}$ are spaced of at least $2a^n$. Indeed two distinct values of $A_{t,n}$ are always spaced by a number $d = 2\sum_{i=0}^{n} \epsilon_i a^i$ where for $i = 0, \ldots, n$, $\epsilon_i \in \{-1, 0, 1\}$. As $l = \min\{i; 0 \leq i \leq n, \epsilon_i \neq 0\}$ exists and $\epsilon_l = 1$, we have $d \geq 2a^n$.

2. We have $\mathbb{P}(a < |X_t| \leq 2) \geq 1/4$. Indeed $\mathbb{P}(a < |X_t| \leq 2) \geq> 0$, $X_t \geq 1 + a - \sum_{i\geq 2} a^i$ for $a \in (0, 1/2]$ if $\xi_t = \xi_{t-1} = 1$. Moreover as $X \leq 1/(1-a) \leq 2$ for $a \in (0, 1/2]$.

3. For $B = (-a, a)$ we have $\mathbb{P}(X_t \in B) > 0$. For this, observe first that $a \in]\frac{3-\sqrt{5}}{2}, 1/2]$ implies $1 - a - a^2 - a^3 - \ldots < a$; thus for $n_0 \geq 2$ large enough we get $1 - a - \ldots - a^{n_0} + \sum_{k\geq n_0+1} a^k < a$.

If $\xi_{t-i} = 1$ for $i \neq 1$ with $0 \leq i \leq n_0$, and $\xi_{t-1} = -1$, we have $0 \leq X_t \leq 1 - a - \ldots - a^{n_0} + \sum_{k\geq n_0+1} a^k < a$. Thus $\mathbb{P}(|X_t| < a) \geq 2^{-n_0-1}$. Now if $w_1, \ldots, w_k$ denote the values of $A_{t,n}$, we set $A = \cup_{i=1}^{k}]w_i - a^{n+2}, w_i + a^{n+2}[$. Using the decomposition we infer that $X_t \in A$if $|X_{t-(n+1)}| < a$ thus $\mathbb{P}(X_t \in A | X_{t-(n+1)} \in B) = 1$.

We prove here that $\mathbb{P}(X_t \in A) < 1$. If $a < |X_{t-(n+1)}| \leq 2$, then X_t writes as $w_i + c$ with $2a^{n+1} \geq |c| > a^{n+2}$. In this case $X_t \notin A$. Indeed $|X_t - w_i| > a^{n+2}$ and if, for example $c > 0$, we use point 1 and the fact that $a \leq 1/2$ to derive: $X_t < w_i + 2a^{n+1} \leq w_{i+1} - a^{n+2}$ provided w_{i+1} exists (else we have obviously $X_t \notin A$). And we obtain $X_t \notin A$ if $c > 0$. It is also the case if $c < 0$ with a similar argument. The result follows from $\mathbb{P}(X_t \in A) = \mathbb{P}(X_t \in A \cap |X_{t-(n+1)}| \leq a) \leq \mathbb{P}(|X_0| \leq a) < 1$. Moreover it is clear that $P(|X_{t-(n+1)}|^2 a) \neq 0$.(Doukhan et al. (2009))

2.1.2 Weak dependence

Doukhan and Louhichi (1999) aim at defining weak dependence coefficients which makes explicit the asymptotic independence between 'past' and 'future'; this means that the 'past' is progressively forgotten. In terms of the initial time series, 'past' and 'future' are elementary events given through finite dimensional marginals. Roughly speaking, for convenient functions f and g, one shall assume that

$$\mathrm{Cov}\,(f(\text{'past'}), g(\text{'future'}))$$

is small when the distance between the 'past' and the 'future' is sufficiently large. Such inequalities are significant only if the distance between indices of the initial time series in the 'past' and the 'future' terms grows to infinity:

$$|\mathrm{Cov}\,(f(P), g(F))| \leq \psi(u, v, \mathrm{Lip}\, f, \mathrm{Lip}\, g)\epsilon(r).$$

Consider $(X_t)_{t\in\mathbb{Z}}$ a process with values in a Polish space $(E, \|\cdot\|)$. $\|\cdot\|_m$ denotes the usual $\mathbb{L}^m$-norm, i.e., $\|X\|_m^m = \mathbb{E}\|X\|^m$ for $m \geq 1$ for every E-valued random variable X. We define the Lipschitz constant in order to distinct functions ψ. For $h : E^u \to \mathbb{R}$,

$$\mathrm{Lip}\, h = \sup_{(y_1,\ldots,y_u)\neq(x_1,\ldots,x_u)\in E^u} \frac{|h(y_1,\ldots,y_u) - h(x_1,\ldots,x_u)|}{\|y_1 - x_1\| + \ldots + \|y_u - x_u\|}.$$

Definition 2.1.1. *Let $(X_t)_{t\in\mathbb{Z}}$ be a process with values in E. Let $\Gamma(u,v,r)$ be the set of (i,j) in $\mathbb{Z}^u \times \mathbb{Z}^v$ such that $i_1 \leq \ldots \leq i_u \leq i_u + r \leq j_1 \ldots \leq j_v$. For some classes of functions $\mathcal{F}_u$ from E^u to $\mathbb{R}$ and $\mathcal{G}_v$ from E^v to $\mathbb{R}$, if ψ is some function from $\mathcal{F}\times\mathcal{G}\times\mathbb{R}^2$ to $\mathbb{R}^+$, the dependence coefficient is defined by*

$$\epsilon(r) = \sup_{u,v} \sup_{(i,j)\in\Gamma(u,v,r)} \sup_{f\in\mathcal{F}_u, g\in\mathcal{G}_v} \frac{|Cov\,(f(X_{i_1},\ldots,X_{i_u}), g(X_{j_1},\ldots,X_{j_v}))\,|}{\psi(f,g,u,v)}.$$

X_t is called (ϵ,ψ)-weakly dependent process if the sequence $\epsilon(r) \to_{r\to\infty} 0$.

Remark that in the previous definition:

a) r always denotes the gap in time between 'past' and 'future'.

b) the sequence ϵ depends both on the class $\mathcal{F},\mathcal{G}$ and on the function ψ.

Assume that $\mathcal{F}_u$ are the set of functions bounded by 1 (resp. $\mathcal{G}_v$). Then the weak dependence coefficients correspond to:

$$\begin{array}{lll}
\psi & = u\mathrm{Lip}\, f + v\mathrm{Lip}\, g & \text{then denote } \epsilon(r) = \eta(r)\\
& = v\mathrm{Lip}\, g & \text{then denote } \epsilon(r) = \theta(r)\\
& = uv\mathrm{Lip}\, f \cdot \mathrm{Lip}\, g & \text{then denote } \epsilon(r) = \kappa(r)\\
& = u\mathrm{Lip}\, f + v\mathrm{Lip}\, g + uv\mathrm{Lip}\, f \cdot \mathrm{Lip}\, g & \text{then denote } \epsilon(r) = \lambda(r)\\
& = u\mathrm{Lip}\, f + v\mathrm{Lip}\, g + uv\mathrm{Lip}\, f \cdot \mathrm{Lip}\, g + u + v & \text{then denote } \epsilon(r) = \omega(r)
\end{array}$$

Remark 2.1.1. *The coefficients η,κ,λ, and ω are non-causal coefficients when $\mathcal{F}_u = \mathcal{G}_u$ and ψ is symmetric. In this situations where both $\mathcal{F}_u$ and $\mathcal{G}_u$ are spaces of regular functions, we say that we are in the non causal case. In the case where the sequence $(X_t)_{t\in\mathbb{Z}}$ is an adapted process with respect to some increasing filtration $(\mathcal{M}_i)_{i\in\mathbb{Z}}$, it is often more suitable to work without assuming any regularity conditions on $\mathcal{F}_u$. In that case $\mathcal{G}_u$ is some space of regular functions and $\mathcal{F}_u \neq \mathcal{G}_u$. This last case is called the causal case.*

An important point in the previous definition is its heredity through appropriate images as is the case for mixing conditions. As well as mixing coefficients, these coefficients also have some hereditary properties.

Proposition 2.1.2 (Bardet et al. (2007))**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a sequence of $\mathbb{R}^k$-valued random variables. Let $p > 1$. We assume that there exists some constant $C > 0$ such that $\max_{0\le i\le k} \|X_i\|_p \le C$. Let h be a function from $\mathbb{R}^k$ to $\mathbb{R}$ such that $h(0) = 0$ and for $x, y \in \mathbb{R}^k$, there exists a in $[1, p[$ and $c > 0$ such that*

$$|h(x) - h(y)| \le c|x - y|(1 + |x|^{a-1} + |y|^{a-1})$$

. We define the sequence $(Y_t)_{t\in\mathbb{Z}}$ by $Y_t = h(X_t)$, then,

- *if $(X_t)_{t\in\mathbb{Z}}$ is θ-weakly dependent, then $(Y_t)_{t\in\mathbb{Z}}$ too, $\theta_Y(r) = \mathcal{O}\left(\theta(r)^{\frac{p-a}{p-1}}\right)$;*
- *if $(X_t)_{t\in\mathbb{Z}}$ is η-weakly dependent, so is $(Y_t)_{t\in\mathbb{Z}}$ and $\eta_Y(r) = \mathcal{O}\left(\eta(r)^{\frac{p-a}{p-1}}\right)$;*
- *if $(X_t)_{t\in\mathbb{Z}}$ is λ-weakly dependent, $(Y_t)_{t\in\mathbb{Z}}$ also $\lambda_Y(r) = \mathcal{O}\left(\lambda(r)^{\frac{p-a}{p+a-2}}\right)$.*

Example 2.1.3. *The function $h(x) = x^2$ satisfies the previous assumptions. This condition is satisfied by polynomials with degree a.*

Let $\mathcal{F}_u$ be the class of bounded functions from E^u to $\mathbb{R}$, and let $\mathcal{G}_u$ be the class of functions from E^u to $\mathbb{R}$ which are Lipschitz. We assume that the variables X_i are $\mathbb{L}^1$-integrable. We shall see that the θ causal coefficient defined above belongs to a more general class of dependence coefficients defined through conditional expectations with respect to the filtration $\sigma(X_j, j \le i)$.

Definition 2.1.2. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $\mathcal{M}$ be a σ-algebra of $\mathcal{A}$. Let E be s Polish space. For any $\mathbb{L}^p$-integrable random variable X with values in E, we define*

$$\theta_p(\mathcal{M}, X) = \sup\{\|\mathbb{E}(g(X)|\mathcal{M}) - \mathbb{E}(g(X))\|_p, \ \ Lip\, g \le 1\}.$$

and then if $(X_i)_{i\in\mathbb{Z}}$ is an $\mathbb{L}^p$-sequence, and $(\mathcal{M}_k)_{k\in\mathbb{Z}}$ are σ-algebras ($\sigma(X_j, j \le k)$).

$$\theta_{p,k}(r) = \max_{s\le k} \frac{1}{s} \sup_{i+r\le j_1\le\ldots\le j_s} \theta_p\left(\mathcal{M}_i, (X_{j_1}, \ldots, X_{j_s})\right).$$

The two preceding definitions are coherent as proved in Dedecker et al. (2007), $\theta(r) = \theta_{1,\infty}(r)$.

Remark 2.1.2. *It is clear that if X is a θ-weakly dependent process it is also a λ-weakly dependent process. Then main reasons for considering a distinction between causal and non causal time series are*

a) the θ-weak dependence is more easily relied to the strong mixing property;

b) some models or properties require different conditions on the convergence rate of $(\theta(r))$ than of $(\lambda(r))$.

We now define τ and γ causal coefficients.

Definition 2.1.3. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $\mathcal{M}$ be a σ-algebra of $\mathcal{A}$. Let E be s Polish space and $p \in [1, \infty]$. For any $\mathbb{L}^p$-integrable random variable X with values in E, we define*

- *τ coefficients:*

$$\tau_p(\mathcal{M}, X) = \| \sup_{Lip\, g \leq 1} \left\{ \int g(X) \mathbb{P}_{X|\mathcal{M}} \mathrm{d}x - \int g(X) \mathbb{P}_X \mathrm{d}x \right\} \|_p.$$

and we clear have

$$\theta_p(\mathcal{M}, X) \leq \tau_p(\mathcal{M}, X)$$

now let $(X_i)_{i \in \mathbb{Z}}$ be a $\mathbb{L}^p$ integrable random sequence. The coefficient $\tau_{p,k}(r)$ are defined as follow:

$$\tau_{p,k}(r) = \max_{s \leq k} \frac{1}{s} \sup_{i+r \leq j_1 \leq \ldots \leq j_s} \tau_p \left(\mathcal{M}_i, (X_{j_1}, \ldots, X_{j_s}) \right).$$

- *γ coefficients (projective measure)*

$$\gamma_p(\mathcal{M}, X) = \|\mathbb{E}(X|\mathcal{M}) - \mathbb{E}(X)\| p \leq \theta_p(\mathcal{M}, X)$$

and

$$\gamma_p(r) = \sup_{i \in \mathbb{Z}} \gamma_p \left(\mathcal{M}_i, X_{i+r} \right).$$

Those coefficients are defined in Gordin (1969), these coefficients are used in order to derive various limit theorems in Mc Leish (1975,a).

2.1.3 Physique dependence measure

In this section, we introduce another look at the fundamental issue of dependence. By interpreting causal Bernoulli shifts as physical systems, Wu (2005) introduce physical and predictive dependence measures quantify the degree of dependence of outputs X_t on inputs ε_t in physical systems. Consider the causal Bernoulli shift

$$X_t = H(\xi_t, \xi_{t-1}, \ldots)$$

where ξ_t, $t \in \mathbb{Z}$ are i.i.d random variables and H is a measurable function. In view as physical system, $\xi_t, \xi_{t-1}, \ldots$ are inputs and H is a filter or a transform. X_i shall be the output. Applying the idea of coupling, they introduce dependence coefficient by measuring the degree of dependence of outputs on inputs. Let (ξ_i') by an iid copy of (ξ_i). Hence $\xi_i', \xi_j, i, j \in \mathbb{Z}$, are i.i.d.

Definition 2.1.4. *Let the shift process $\mathcal{F}_i = (\xi_i, \xi_{i-1}, \ldots)$. Denote X_j^* be a coupled version of X_j in the latter beging replaced by ξ_0':*

$$X_j^* = H(\mathcal{F}_j^*),\ \mathcal{F}_j^* = (\xi_j, \xi_{j-1}, \ldots, \xi_1, \xi_0', \xi_{-1}, \ldots).$$

For $j \in \mathbb{Z}$, define the the projection operator

$$\mathcal{P}_j(X) = \mathbb{E}(X|\mathcal{F}_j) - \mathbb{E}(X|\mathcal{F}_{j-1})$$

- *Functional or physical dependence measure. Let $X_i \in \mathbb{L}^p$, $p > 0$,*

$$\delta_p(j) = \|X_j - X_j^*\|_p$$

- *Predictive dependence measure. Let $X_i \in \mathbb{L}^p$, $p \geq 1$*

$$\theta_p(i) = \|\mathcal{P}_0 X_i\|_p$$

- *p-stability. The process (X_t) is said to be p-stable if*

$$\Delta_p := \sum_{j=0}^{\infty} \delta_p(j) < \infty.$$

  We say that it is weakly p-stable if

$$\Omega_p := \sum_{j=0}^{\infty} \theta_p(j) < \infty.$$

Limit theorems with those dependence measures have been established and are often optimal or nearly optimal. Those dependence measures provide a simple way for a large-sample theory for stationary causal processes and they are directly related to the underlying data-generating mechanism H. Examples as linear processes and Volterra processes, a polynomial-type nonlinear process, nonlinear time series ...

2.2 Models

2.2.1 Bernoulli shifts

Now we consider the weak dependence structure to the class of Bernoulli shifts.

Definition 2.2.1. *Let ξ_i, $i \in \mathbb{Z}$, be independent and identically distributed random variables and H a measurable function defined on $\mathbb{R}^{\mathbb{Z}}$. A Bernoulli shift is a sequence $(X_t)_{t\in\mathbb{Z}}$ defined by*

$$X_t = H((\xi_{t-j})_{j\in\mathbb{Z}}),$$

where, more precisely, H in $\mathbb{L}^m(\mu)$ for some $m > 0$, with μ the distribution of $(\xi_t)_{t\in\mathbb{Z}}$.

This way of constructing stationary sequence is very natural. A simple case of infinitely dependent Bernoulli shift is the moving average process, writes $X_t = \sum_{j=-\infty}^{\infty} a_j \xi_{t-j}$.

Proposition 2.2.1 (Doukhan and Louhichi (1999))**.** *The process $(X_t)_{t\in\mathbb{Z}}$ is η-weak dependent with $\eta(r) = 2\delta_{[r/2]}^{m\wedge 1}$ if*

$$\mathbb{E}|H(\xi_j, j \in \mathbb{Z}) - H(\xi_j \mathbb{1}_{|j|<r}, j \in \mathbb{Z})| \leq \delta_r \downarrow 0 (r \uparrow \infty) \tag{2.2}$$

If $H(\xi_j, j \in \mathbb{Z})$ does not depend on ξ_j with $j < 0$, then it is causal and θ-dependent holds with $\theta(r) = \delta_r^{m\wedge 1}$.

In fact, the sequences $(\delta_k)_k$ are related to the modulus of uniforms continuity of H. It is evaluated under regularity conditions on the function H; e.g. if

$$|H(u_i; i \in \mathbb{Z}) - H(v_i; i \in \mathbb{Z})| \leq \sum_{i\in\mathbb{Z}} a_i |u_i - v_i|^b$$

for some $0 < b \leq 1$ and for positive constants $(a_i)_{i\in\mathbb{Z}}$ fulfilling $\sum_{i\in\mathbb{Z}} a_i < \infty$. If the sequence $(\xi_i)_{i\in\mathbb{Z}}$ has finite bth-order moment, then

$$\delta_k \leq \sum_{|i|\geq k} a_i E|\xi_i|^b.$$

Notice finally that most of models used in statistics are such processes. Examples of such situations follow:

- Example (2.1.1), the example of the non mixing stationary Markovian chain with i.i.d Binomial innovations,

  $$X_t = \frac{1}{2}(X_{t-1} + \xi_t)$$

  satisfies $\delta_r = O(2^{-r})$; its marginal distribution is uniform on $[0, 1]$.

- *Nonparametric AR model*

  The real-valued functional autoregressive model

  $$X_t = r(X_{t-1}) + \xi_t \text{ with } r : \mathbb{R} \to \mathbb{R}.$$

  If $|r(u)-r(u')| \leq c|u-u'|$ for some $0 \leq c \leq 1$ and for all $u, u' \in \mathbb{R}$, and if the i.i.d. innovation process $(\xi_t)_{t\in\mathbb{Z}}$ satisfies $E\|X_0\| < \infty$, then θ-dependence holds with $\theta(r) = \delta_r = C \cdot c^r$ for some constant $C > 0$.

2.2.2 Models with a Markovian representation

Let $(X_t)_{t\in\mathbb{N}}$ be sequence of random variables with values in a Banach space $(\mathbb{B}, \|\cdot\|)$. Let $(\xi_t)_{t\in\mathbb{N}}$ be a sequence of independent r.v.s and F be a measurable function. Assume that X_t satisfies the recurrence equation

$$X_t = F(X_{t-1}, \xi_t).$$

The initial distribution X_0 is supposed to be independent of the sequence $(\xi_i)_{i\in\mathbb{N}}$. Assume that, the function F satisfies

$$\begin{cases} \mathbb{E}\|F(0, \xi_1)\|^a < \infty \\ \mathbb{E}\|F(x, \xi_1) - F(y, \xi_1)\|^a \leq \alpha^a \|x - y\|^a \end{cases} \tag{2.3}$$

for some $a \geq 1$ and $0 \leq \alpha < 1$. It is known by Duflo (1996) that the Markov chain $(X_i)_{i\in\mathbb{N}}$ has a stationary law μ with finite moment of order a. We suppose

that μ is the distribution of X_0(i.e the Markov chain is stationary). If moreover condition (2.3) is satisfied then the Markov chain, if $\tilde{X}_0$ is independent of X_0 and distributed as X_0, previous defined is weakly dependent and

$$\theta_{p,\infty}(r) \leq \tau_{p,\infty}(r) \leq \alpha^r \|\tilde{X}_0 - X_0\|_p$$

(see Doukhan and Louhichi (1999)).

Remark that the stationary iterative markov models $X_t = F(X_{t-1}, \xi_t)$ can be represented as Bernoulli shifts if condition (2.3) holds, when X_t and ξ_t take values in Euclidean space.

2.2.3 Linear process

If X is a ARMA(p, q) process or, more generally, linear process such that

$$X_t = \sum_{j=0}^{\infty} a_j \xi_{t-j}$$

for $t \in \mathbb{Z}$, with $a_j = O(|j|^{-\mu})$ with $\mu > 1/2$.

A first choice is

$$\delta_r = \mathbb{E}|\xi_0| \sum_{k>r} |a_k|$$

for the linear process with i.i.d innovations such that $\mathbb{E}|\xi_0| < \infty$.

For centered and $\mathbb{L}^2$ innovations, another choice is

$$\delta_r = \sqrt{\mathbb{E}|\xi_0|^2 \sum_{k>r} |a_k|^2}.$$

Thus X is a θ- (respectively, λ-) weakly dependent process with

$$\theta(r) = \lambda(r) = O\left(\frac{1}{r^{\mu-1/2}}\right)$$

(see Doukhan and Lang (2002)). It is also possible to deduce λ-weak dependence properties for X if the innovation process is itself λ-weakly dependent (Doukhan and Wintenberger (2008)).

2.2.4 Chaotic expansion

We study chaotic expansions associated with the discrete chaos generated by the sequence $(\xi_t)_{t\in\mathbb{Z}}$. In a condensed formulation we write

$$F(x) = \sum_{k=0}^{\infty} F_k(x)$$

where $F_k(x)$ denote the kth order chaos contribution and $F_0(x) = a_0^{(0)}$ is only a centring constant and

$$F_k(x) = \sum_{j_1=-\infty}^{\infty} \sum_{j_2=-\infty}^{\infty} \dots \sum_{j_k=-\infty}^{\infty} a^{(k)}_{j_1,\dots,j_k} x_{j_1} \times x_{j_2} \times \dots \times x_{j_k}.$$

It can be written in the vectorial notation $F_k(x) = \sum_{j\in\mathbb{Z}^k} a_j^{(k)} x_j$.

An example is a Volterra stationary process defined through a convergent Voleterra expansion

$$X_t = v_0 + \sum_{k=1}^{\infty} V_{k;t} \;\; V_{k;t} = \sum_{-\infty<j_1<\dots<j_k<\infty} a^{(k)}_{j_1,\dots,j_k} - \xi_{t-j_1} \dots \xi_{t-k},$$

where v_0 denotes a constant and $(a_j^{(k)})_{j\in\mathbb{Z}^k} = (a^{(k)}_{j_1,\dots,j_k})_{(j_1,\dots,j_k)\in\mathbb{Z}^k}$ are real number for each $k > 0$. This expression converges in $\mathbb{L}^m$ for $m \geq 1$, provided that $\mathbb{E}|\xi_0|^m < \infty$ and $\sum_{k=0}^{\infty} \sum_{j\in\mathbb{Z}^k} |a_j^{(k)}| < \infty$. Those models are η-dependent since (2.2) is satisfied, δ_r corresponding to the tail of the previous series

$$\delta_r = \sum_{k=0}^{\infty} \left\{ \sum_{j\in\mathbb{Z}^k;\|j\|_\infty>r} |a_j^{(k)}| \mathbb{E}|\xi_0|^k \right\} < \infty.$$

One more example is the simple bilinear process with the recurrence equation

$$X_t = aX_{t-1} + bX_{t-1}\xi_{t-1} + \xi_t.$$

Such processes are associated with the chaotic representation in

$$F(x) = \sum_{j=1}^{\infty} x_j \prod_{s=0}^{j-1} (a + bx_s), \;\; x \in \mathbb{R}^{\mathbb{Z}}$$

If $c = \mathbb{E}|a + b\xi_0| < 1$ then $\delta_r = \theta_r = c^r(r+1)/(c-1)$ has a geometric decay rate.

2.2.5 LARCH(∞) models

We mention LARCH(∞) models from Doukhan et al. (2006). Let $(\xi_t)_{t\in\mathbb{Z}}$ be an i.i.d sequence of random $d \times D$-matrices, $(A_j)_{j\in\mathbb{N}^*}$ be a sequence of $D \times d$ matrices, and a be a vector in $\mathbb{R}^D$. Conditionally heteroscedastic models can be expressed in terms of a vector valued LARCH(∞) model, which is a solution of the recurrence equation

$$X_t = \xi_t \left(a + \sum_{j=1}^{\infty} A_j X_{t-j} \right)$$

Such models are proved to have a stationary representation with the chaotic expansion

$$X_t = \xi_t \left(a + \sum_{k=1}^{\infty} \sum_{j_1,\dots,j_k \geq 1} A_{j_1} \xi_{t-j_1} A_{j_2} \dots A_{j_k} \xi_{t-j_1-j_2-\dots-j_k} a \right) \tag{2.4}$$

If $\phi = \|\xi_0\|_m \sum_j \|a_j\| < 1$, there exists a solution of previous LARCH models for some $m \geq 1$, and it's given as (2.4). This solution has been proved weakly-dependent with $\theta(r) \leq \|X_r - \tilde{X}_r\|_1$ and $\tau_{m,\infty}(r) \leq \|X_r - \tilde{X}_r\|_m$ where

$$\|X_r - \tilde{X}_r\|_m \leq \|\xi_0\|_m \left(\|\xi_0\|_m \sum_{j<t} j\phi^{j-1} A\left(\frac{t}{j}\right) + \frac{\phi^r}{1-\phi} \right)$$

with $A(s) = \sum_{j\geq s} \|a_j\|$. Moreover for some constants C, C' and b,

$$\theta(r) \leq \begin{cases} C' \frac{(\log(r))^{b\vee 1}}{r^b}, & \text{under Riemaniann decay } A(s) \leq Cs^{(-b)}, \\ C'(q \vee \phi)^{\sqrt{r}}, & \text{under geometric decay } A(s) \leq Cq^s. \end{cases}$$

Such LARCH(∞) models include a large variety of models, as

- Bilinear models

$$X_t = \zeta_t \left(a + \sum_{j=1}^{\infty} \alpha_j X_{t-j} \right) + \beta + \sum_{j=1}^{\infty} \beta_j X_{t-j}$$

where the variables are real valued and ζ is the innovation. For this, we set $\xi_t = \begin{pmatrix} \zeta_t \\ 1 \end{pmatrix}$, $a = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ and $A_j = \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix}$.

- ARCH(∞) processes,

$$\begin{cases} r_t = \sigma_t \xi_t \\ \sigma_t^2 = \beta_0 + \sum_{j=1}^{\infty} \beta_j \sigma_{t-j}^2 \end{cases},$$

We set $\xi_t = (\xi_t 1)$, $a = \begin{pmatrix} \kappa\beta_0 \\ \lambda_1\beta \end{pmatrix}$, $A_j = \begin{pmatrix} \kappa\beta_j \\ \lambda_1\beta_j \end{pmatrix}$ with $\lambda_1 = \mathbb{E}(\xi_0^2)$ and $\kappa^2 = \text{Var}\,(\xi_0^2)$.

- GARCH(p, q) process,

$$\begin{cases} r_t = \sigma_t \xi_t \\ \sigma_t^2 = \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 + \gamma + \sum_{j=1}^{q} \gamma_j r_{t-j}^2 \end{cases},$$

where $\gamma > 0$, $\gamma_i \geq 0$, $\beta_i \geq 0$, and the variables ξ_t are centered at expectation.

2.2.6 Models with infinite memory

Let $(\xi_t)_{t\in\mathbb{Z}}$ be i.i.d, and $F : (\mathbb{R}^d)^{\mathbb{N}} \times \mathbb{R}^D \to \mathbb{R}^d$, we introduce a chain with infinite memory as the stationary solution of the equation

$$X_t = F(X_{t-1}, X_{t-2}, X_{t-3}, \ldots; \xi_t).$$

Assume, for some $m \geq 1$, that $A = \|F(0, 0, 0, \ldots; \xi_t)\|_m < \infty$ and

$$\|F(x_1, x_2, x_3, \ldots; \xi_t) - F(y_1, y_2, y_3, \ldots; \xi_t)\|_m \leq \sum_{j=1}^{\infty} a_j \|x_j - y_j\|.$$

where $(a_j)_{j\geq 1}$ is a sequence of non-negative real number such that

$$a = \sum_{j=1}^{\infty} a_j < 1.$$

Then existence of the model holds in $\mathbb{L}^m$, as well as its stationarity and its weak dependence with,

$$\theta(r) \leq C \inf_{N>0} \left(\sum_{j\geq N} a_j + e^{-\alpha r/N} \right)$$

if $e^{-\alpha} = \sum_{j=1}^{\infty} a_j$, or

$$\lambda(r) = \inf_{p \geq 1} \{a^{r/p} + \sum_{|j|>p} a_j\}.$$

(see Doukhan and Wintenberger (2008))

Chains with infinite memory can also be represented as causal Bernoulli shifts $X_t = H(\xi_t, \xi_{t-1}, \xi_{t-2}, \ldots)$, and then conditions on H gave weak dependence properties and asymptotic results. But several Bernoulli shifts, such as Volterra series, may not fit the parsimony criterion and the function H may be non-explicit.

2.2.7 Gaussian and associated processes

Definition 2.2.2. *The sequence $(Z_t)_{t\in\mathbb{Z}}$ is associated, if for all coordinatewise increasing real-valued functions h and k,*

$$Cov(h(Z_t, t \in A), k(Z_t, t \in B)) \geq 0$$

for all finite subsets A and B of $\mathbb{Z}$.

Gaussian or associated $\mathbb{L}^2$-processes are weakly dependent if

$$\kappa(r) = \mathcal{O}\left(\sup_{i \geq r} |\mathrm{Cov}\,(X_0, X_i)|\right) \to_{r\to\infty} 0.$$

then X is a λ-weakly dependent process such that $\lambda_r = \mathcal{O}\left(\sup_{i\geq r} |\mathrm{Cov}\,(X_0, X_i)|\right)$. See Doukhan and Louhichi (1999) for more details.

Chapter 3

Dependence of Integer Valued Time Series

3.1 Introduction

Over the last fifty years or so, various dependence conditions have emerged in the literature, as a result of the notion of mixing introduced by Rosenblatt (see Rosenblatt (1985) for more). Mixing notions have been applied to numerous dependence type problems; especially in the context of time series and their financial applications they were applied on proving limit theorems which enable valid asymptotic inference; see Doukhan (1994), Rio (2000) and Bradley (2007) for further examples. However, for some models encountered frequently in applications, strong mixing conditions are not satisfied. Prominent examples of such models are the celebrated AR(1) non-mixing model of Andrews (1984) and the LARCH(1) model considered by Doukhan et al. (2009). These types of problems motivated Doukhan and Louhichi (1999) to introduce more flexible dependence conditions to accommodate larger classes of time series models. The main notion introduced is that of weak dependence; the topic is studied extensively in the recent monograph by Dedecker et al. (2007) which includes numerous examples of weakly dependent processes.

The goal of this section is to investigate the relationship between mixing and weak dependence for integer valued time series models. In recent years, there is an emerging literature on the topic of modeling and inference for count time

series, see Kedem and Fokianos (2002), Doukhan et al. (2006), Drost et al. (2008) Fokianos et al. (2009), Fokianos and Tjøstheim (2011), Franke (2010) and Neumann (2011) for integer autoregressive models and for generalized autoregressive models, among other references. We will focus on such models but we point out that other families might be considered as well; see Coupier et al. (2006) for the case of a general process with two values. The objective is to relate mixing and weak dependence conditions for such integer valued count time series models. Due to the fact that the σ-algebras generated by discrete sets are quite small, we prove that the coefficients obtained from the mixing world often coincide to those introduced under weak dependence. The case of Markov processes is of a particular interest in our investigation. Several examples of integer autoregressive models are discussed in detail. In particular, we will prove conditions which existing models should satisfy so that they are weakly dependent. In this way, we offer several theoretical tools for estimation and inference about integer autoregressive processes.

Theorem 3.2.1 gives conditions for the existence and stationarity of a rich class of time series models; see Doukhan and Wintenberger (2008). Section 3.3 contains several new results for integer valued time series models; in particular it links the various coefficients of dependence with special attention to Markov chains. The section 3.4 contains several examples and discusses conditions for their weak dependence by utilizing suitably Theorem 3.2.1.

3.2 Generalities

For the Euclidean space $\mathbb{R}^d$ equipped with some norm $\|\cdot\|$, define the space $\Lambda_1\left(\mathbb{R}^d\right)$ by the set of functions $h : \mathbb{R}^d \to \mathbb{R}$ such that $\operatorname{Lip} h \leq 1$. Furthermore, let us denote by $\|h\|_\infty = \sup_{x\in\mathbb{R}^d} |h(x)|$.

We will be working with the notion of τ–dependence as introduced by Dedecker and Prieur (2004); this notion seems to be appropriate for integer valued time series models. To be more specific, let $(\Omega, \mathcal{G}, \mathrm{P})$ be a probability space and suppose that $\mathcal{M}$ is a σ-algebra of $\mathcal{G}$. We denote by $\mathbb{L}^m(\Omega, \mathcal{G}, \mathrm{P})$ the class of measurable functions $g(\cdot)$, such that $\|g\|_m = (\int_\Omega |g(x)|^m \mathrm{dP}(x))^{1/m} < \infty$. Let X be a random variable on $(\Omega, \mathcal{G}, \mathrm{P})$ with values in $\mathbb{R}^d$. Assume that $\|X\|_1 < \infty$ and define the

coefficient τ as

$$\tau(\mathcal{M}, X) = \left\| \sup \left\{ \left| \int f(x) \mathrm{P}_{X|\mathcal{M}}(\mathrm{d}x) - \int f(x) \mathrm{P}_X(\mathrm{d}x) \right| \Big/ f \in \Lambda_1\left(\mathbb{R}^d\right) \right\} \right\|_1 .$$

An easy way to bound this coefficient is based on a coupling argument; it can be shown that

$$\tau(\mathcal{M}, X) \leq \|X - Y\|_1,$$

for any random variable Y with the same distribution as X and independent of $\mathcal{M}$, see Dedecker and Prieur (2004). As those authors, we assume that the probability space $(\Omega, \mathcal{G}, \mathrm{P})$ is rich enough to define independent sequences of random variables. This implies that there exists a random variable X^* such that

$$\tau(\mathcal{M}, X) = \|X - X^*\|_1.$$

Using the definition of τ, the dependence between the past of the sequence $(X_t)_{t\in\mathbb{Z}}$ and its future k-tuples may be assessed as follows. For two k-tuples $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$, consider the norm $\|x - y\| = \|x_1 - y_1\| + \cdots + \|x_k - y_k\|$ on $\mathbb{R}^{dk}$, set $\mathcal{M}_p = \sigma(X_t, t \leq p)$ and

$$\begin{aligned} \tau_k(r) &= \max_{1\leq l\leq k} \frac{1}{l} \sup \left\{ \tau(\mathcal{M}_p, (X_{j_1}, \ldots, X_{j_l})) \Big/ p + r \leq j_1 < \cdots < j_l \right\}, \quad (3.1) \\ \tau(r) &= \sup_{k>0} \tau_k(r). \quad (3.2) \end{aligned}$$

Then, we say that the time series $(X_t)_{t\in\mathbb{Z}}$ is τ-weakly dependent when its coefficients $\tau(r)$ tend to 0 as r tends to infinity.

Note that the last condition implies other notions of dependence; the η and θ-weak dependence. Consider numeric functions f and g uniformly bounded by 1 and defined on the sets $(\mathbb{R}^d)^u$ and $(\mathbb{R}^d)^v$ equipped with the following norm.

$$\|(x_1, \ldots, x_u)\| = \|x_1\|_\infty + \cdots + \|x_u\|_\infty, \qquad x_1, \ldots, x_u \in \mathbb{R}^d.$$

where $\|x\|_\infty = \max_{1\leq j\leq d} |x_j|$, for any $x \in \mathbb{R}^d$. Then those coefficients are defined as the least nonnegative numbers $\eta(r)$ and $\theta(r)$ such as

$$\begin{aligned} \left| \mathrm{Cov}\left(f(X_{i_1}, \ldots, X_{i_u}), g(X_{j_1}, \ldots, X_{j_v})\right) \right| &\leq (u\mathrm{Lip}\, f + v\mathrm{Lip}\, g)\eta(r) \\ &\leq v\mathrm{Lip}\, g \cdot \theta(r) \end{aligned}$$

for integers $i_1, \ldots, i_u, j_1, \ldots, j_v$ which satisfy $i_1 \leq \cdots \leq i_u \leq i_u + r \leq j_1 \leq \cdots \leq$

j_v. Note that $\eta(r) \leq \theta(r) \leq \tau(r)$ and the definition of $\eta(r)$ corresponds to the case of non causal models.

The following theorem gives a general result about the decay rate of weak dependence coefficients and improves upon the results obtained by Doukhan and Wintenberger (2008) for infinite order models, which are not, in general, Markov models; see e.g. LARCH(∞) models in Dedecker et al. (2007).

Theorem 3.2.1. *Suppose that $\{X_t, t \in \mathbb{Z}\}$ is a time series which satisfies*

$$X_t = F(X_{t-1}, X_{t-2}, \ldots; \xi_t), \tag{3.3}$$

where $\{\xi_t, t \in \mathbb{Z}\}$ is an i.i.d sequence. Suppose that the function $F(\cdot)$ satisfies the following conditions:

$$\begin{aligned} \|F(0, \xi_0)\|_m &< \infty \\ \|F(x; \xi_0) - F(x'; \xi_0)\|_m &\leq \sum_{l=1}^{\infty} \alpha_l \|x_l - x_l'\|, \end{aligned}$$

where $x = (x_i)_{i \geq 1}$, $x' = (x_i')_{i \geq 1}$ belong to $\mathbb{R}^\infty$ and $(\alpha_l)_{l \geq 1}$ a sequence of positive real numbers with $\alpha = \sum_l \alpha_l$. If $\alpha < 1$, then there exists a unique causal stationary solution X which satisfies equation (3.3) such that $\|X_0\|_m < \infty$. Moreover, $\{X_t, t \in \mathbb{Z}\}$ is both η and τ weakly dependent process with corresponding coefficients.

$$\tau(r) \leq \frac{2\|F(0, \xi_0)\|_1}{1 - \alpha} \inf_{1 \leq u \leq r} \left\{ \alpha^{\frac{r}{u}} + \frac{1}{1 - \alpha} \sum_{k=u+1}^{\infty} \alpha_k \right\}$$

Analogous result holds true for the η-coefficients. In particular, for Markov models, we obtain that the sequence $\tau(r)$ decays exponentially fast.

Following either Doukhan (1994) or Rio (2000), recall that for integers $1 \leq u, v \leq \infty$, the strong mixing coefficient is defined by

$$\alpha_{u,v}(r) = \sup |\mathbb{P}(U \cap V) - \mathbb{P}(U)\mathbb{P}(V)|, \tag{3.4}$$

$$\alpha(r) = \alpha_{\infty,\infty}(r), \tag{3.5}$$

whereas the absolute regularity mixing coefficient is given by

$$\beta_{u,v}(r) = \sup \sum_{i,j} |\mathbb{P}(U_i \cap V_j) - \mathbb{P}(U_i)\mathbb{P}(V_j)|, \tag{3.6}$$

$$\beta(r) = \beta_{\infty,\infty}(r). \tag{3.7}$$

In all the above displays, the suprema is taken over $U \in \mathcal{U}$, and $V \in \mathcal{V}$ or respectively for measurable partitions $U_i \in \mathcal{U}$, and $V_j \in \mathcal{V}$ of Ω; note that $\mathcal{U} = \sigma(X_{i_1}, \ldots, X_{i_u})$, and $\mathcal{V} = \sigma(X_{j_1}, \ldots, X_{j_v})$ for integers $i_1 \leq \cdots \leq i_u \leq i_u + r \leq j_1 \leq \cdots \leq j_v$; the suprema first runs over all such integers and second over the sigma fields $\mathcal{U}, \mathcal{V}$.

3.3 Dependence of integer valued time series

We investigate the relation between mixing and weak dependence for integer valued time series. We first need the following definition.

Definition 3.3.1. *For each $d \geq 1$ we denote by $\|\cdot\|_\infty$ the uniform norm, i.e. $\|(u_1, \ldots, u_d)\| = \max_{1 \leq j \leq d} |u_j|$ on $\mathbb{R}^d$. A set $\mathbb{G}$ will be called discrete if $\mathbb{G} \subset \mathbb{R}^d$ for some $d \geq 1$ and its elements satisfy*

$$D = \inf_{x \neq x', x, x' \in \mathbb{G}} \|x - x'\|_\infty > 0$$

Note that if $\mathbb{G} = \mathbb{Z}^d$, then $D = 1$.

Lemma 3.3.1. *Any real valued function with uniform norm less than 1 defined on $\mathbb{G}$ with $\mathbb{G}$ discrete, is the restriction of a $[-1, 1]-$valued and $\frac{2}{D}$-Lipschitz function.*

Based on the previous lemma, we can link the weak dependence coefficient η to the strong mixing coefficient α.

Proposition 3.3.1. *If $\{X_t, t \in \mathbb{Z}\}$ is an η-weakly dependent integer valued process, then*

$$\alpha_{u,v}(r) \leq \frac{2}{D}(u + v)\eta(r)$$

Note that the same situation applies to the coefficients τ. Hence Lemma 3.3.1 shows again that

Proposition 3.3.2. *If $\{X_t, t \in \mathbb{Z}\}$ is an τ-weakly dependent integer valued process, then*

$$\beta_{\infty,v}(r) \leq \frac{2v}{D}\tau(r)$$

This entails that a different technique should better be chosen to bound from above the usual mixing coefficients α, or β . Towards this goal, either a restriction for the range of the process, similar to Coupier et al. (2006) (where only $\{0,1\}-$discrete valued models are considered) is needed or the restriction to Markov type times series for which memory properties are essential.

We thus specialize the investigation to Markov processes. Assume that $\{X_t, t \in Z\}$ is a $\mathbb{G}^d$-valued stationary Markov process where $\mathbb{G}$ is a discrete set; see Definition 3.3.1. Then employing ideas of Doukhan (1994), we recall that the absolute regularity coefficient has the simple expression

$$\beta(r) = \|\mathbb{P}_{(X_0,X_r)} - \mathbb{P}_{X_0} \otimes \mathbb{P}_{X_r}\|_{TV} = \sup_{\|f\|_\infty \leq 1} |\mathbb{E}(f(X_0, X_r) - f(X_0, X_r^*))|,$$

where X_r^* is a copy of X_r independent of X_0.

Hence the Proposition 3.3.2 together with the fact that for Markov processes $\beta(t) = \beta_{1,1}(t)$, we derive the following result:

Theorem 3.3.1. *Assume that $(X_t)_{t\in\mathbb{Z}}$ is a stationary $p-$Markov chain with values in $\mathbb{G}^d$.*

- *Assume also that this chain is $\eta-$weakly dependent, then this process is absolutely regular.*
- *Moreover, if it is η-weakly dependent and $\|X_0\|_m^m < \infty$ for some $m > 0$, then its absolute regularity coefficient sequence satisfies*

$$\beta(r) \leq 4p2^{\frac{m}{m+d}} D^{-(1+d)} (\|X_0\|_m^m)^{\frac{d}{m+d}} \eta(r)^{\frac{m}{m+d}},$$

  for all $r \in \mathbb{N}$, large enough.
- *If the process is τ-dependent then*

$$\beta(r) \leq \frac{2p}{D}\tau(r)$$

An immediate consequence of the above theorem is that for $d > 1$ yields rates of dependence for d-dimensional Markov integer valued processes. Indeed for a $d-$Markov process $(Z_t)_{t\in\mathbb{Z}}$, setting $X_t = (Z_t, \ldots, Z_{t-d+1})$, the process $(X_t)_{t\in\mathbb{Z}}$ is now a Markov and $\mathbb{G}^d-$valued process.

3.4 Examples

Here we give some examples of integer-valued time series models that are weakly dependent. The great advantage of working with the notion of weak dependence is that the ease of verification of (3.3) which shows that when the function $F(.)$ is Lipschitz, then the process is stationary which possess moments of any order.

3.4.1 Integer autoregressive models of order p

Integer autoregressive processes have been introduced by Al-Osh and Alzaid (1987, 1990) as a convenient way to transfer the usual autoregressive structure to discrete valued time series. The main concept is given by the notion of thinning which is defined by as follows.

Suppose that X is a non–negative integer random variable and let $a \in [0, 1]$. Then, the thinning operator, denoted by $\circ$, is defined as

$$a \circ X = \begin{cases} \sum_{i=1}^{X} Y_i, & \text{if } X > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\{Y_i\}$ is a sequence of independent and identically distributed non-negative integer valued random variables with mean a–independent of X. The sequence $\{Y_i\}$ is termed as a counting series. The most common example is when the counting sequence consists of an iid sequence of Bernoulli random variables with probability of success a.

To carry out the task of identifying the right form of $F(\cdot)$ in (3.3), it is important to use an alternative representation of the thinning operator. More specifically, suppose that $\{U_i, i \geq 1\}$ is a sequence of iid standard uniform random variables. Let $p_a(k) = \mathrm{P}(Y \leq k)$, $k = 0, 1, 2, \ldots$. Then, we can express the random variables Y_i explicitly in terms of the uniform random variables by

$$Y_i(a) = Y(U_i, a) = \sum_{k=0}^{\infty} 1(U_i \leq p_a(k)).$$

The above representation allows for more convenient calculations, as we shall see.

The integer autoregressive process of order p is defined as follows. Suppose that for $i = 1, 2, \ldots, p$, $a_i \in [0, 1)$ and let $\{\xi_t\}$ be a sequence of independent

and identically distributed nonnegative integer valued random variables with $\|\xi\|_r < \infty$. Then, the following process

$$X_t = \sum_{i=1}^{p} a_i \circ X_{t-i} + \xi_t, \tag{3.8}$$

is called integer autoregressive process of order p and is denoted by INAR(p). It should be noted that the counting series used for defining the random variable $a_1 \circ X_{t-1}$ are independent of those involved in the definition of $a_2 \circ X_{t-2}$, and so on. This assumption guarantees that the INAR(p) process has the classical AR(p) correlation structure, see Du and Li (1991). Now using definition (3.8) and the almost sure representation of the thinning operator, we obtain that

$$\begin{aligned} X_t &= \sum_{i=1}^{p} a_i \circ X_{t-i} + \xi_t = F(X_{t-1}, \ldots, X_{t-p}; \xi_t) \\ &= \sum_{i=1}^{p} \sum_{j=1}^{X_{t-i}} Y(U_{t;j} a_i) + \xi_t, \end{aligned}$$

where the error sequence is defined $\xi_t = (\xi_t, V_t)$ with $V_t = (U_{t;j})_{j \geq 1}$. Now, it is easy to verify the conditions that (3.3) has to satisfy. Since $\|\xi\|_r < \infty$ we have that the first condition is satisfied. For the second condition, note that an application of Minkowski shows that

$$\|F(x_1, \ldots, x_p; \xi_0) - F(x_1^{'}, \ldots, x_p^{'}; \xi_0)\|_r \leq \sum_{i=1}^{p} \|Y(a_i)\|_r |x_i - x_i^{'}|.$$

Hence with $\alpha = \sum_{i=1}^{p} \|Y(a_i)\|_r < 1$, the conclusion of Theorem (3.2.1) hold true. In particular, when the counting series is Bernoulli random variables with probability a_i, then we obtain the condition $\sum_{i=1}^{p} a_i < 1$, which is the standard condition for stationarity and ergodicity of the INAR(p) model with Bernoulli counting series, see Du and Li (1991).

Remark 3.4.1. *In order to derive mixing properties of Markov processes one needs irreducibility: this makes a real problem for integer valued models since they belong ṡ. to a null set with respect to Lebesgue measure. In theorem 3.3.1 Lyapounov technique does not apply for the simple Markov models INAR(1):*

$$X_t = a \circ X_{t-1} + \xi_t$$

Even stationarity needs $|a| < 1$. The operator $x \mapsto a \circ x$ is contracting in the mean for this case. Quote that $x \mapsto ax$ is uniformly contracting. Thus Steutel

van Harn operators provide special problems. Indeed, let $a\circ$ *denote the Steutel and van Harn operator based on a counting sequence* $(Y_i)_{i\in\mathbb{N}}$. *We have*

$$\|F(x;\xi_0) - F(x^{'};\ \xi_0)\|_r \leq \max_j |Y_j| \|x - x^{'}\|_r.$$

The $\max_j |Y_j|$ *can be not bounded by* $(0,1)$ *(see Wu and Shao (2004)).*

3.4.2 Integer valued bilinear models

Consider the following bilinear type of INAR model

$$X_t = a_1 \circ X_{t-1} + b_1 \circ (X_{t-1}\xi_{t-1}) + \xi_t,$$

called BINAR(1,1). Then, working analogously as before, we can show that a necessary condition for $\{X_t\}$ to be stationary and ergodic with r-moments is given by

$$\begin{cases} \|\xi_t\|_r < \infty \\ \|Y(a_1)\|_r + \|\epsilon_t\|_r \|Y(b_1)\|_r < 1, \end{cases}$$

see Doukhan et al. (2006), Drost et al. (2008) for more.

3.4.3 Integer valued LARCH models

More generally we can consider integer valued ARCH type models with infinite memory; for instance suppose that

$$X_t = \xi_t \left(a_0 + \sum_{i=1}^{\infty} a_i \circ X_{t-i} \right).$$

Then again, the elementary calculations show that a necessary condition for $\{X_t\}$ to be stationary and ergodic with r-moments is given by

$$\|\xi_t\|_r \sum_{i=1}^{\infty} \|Y(a_i)\|_r < 1.$$

See Latour and Truquet (2008).

3.4.4 Mixed INAR(1) models

Suppose that for all $i \in \{1, 2, \ldots, k\}$, $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$. Then a mixed integer autoregressive model can be considered for modeling when the process changes behavior in different regimes. More precisely, suppose that

$$X_t = \begin{cases} a_1 \circ X_{t-1} + \xi_t, & \text{with probability } p_1, \\ a_2 \circ X_{t-1} + \xi_t, & \text{with probability } p_2, \\ \quad \vdots & \\ a_k \circ X_{t-1} + \xi_t, & \text{with probability } p_k. \end{cases}$$

To examine weak dependence properties of the above model it is convenient to introduce a random variable, say J, which is independent of the counting series and the error terms and such that $\mathrm{P}(J = j) = p_j$, for $j = 1, 2, \ldots, k$. Then, the above process can be rewritten as

$$X_t = \sum_{j=1}^{k} I_{\{J=j\}}(a_j \circ X_{t-1} + \xi_t).$$

Now, it is again simple to show that a stationary and ergodic process $X = (X_t)_{t\in\mathbb{Z}}$ with finite moments of order r satisfies the above model if

$$\begin{cases} \|\xi_t\|_r < \infty \\ \sum_{j=1}^{k} (p_j)^{1/r} \|Y(a_j)\|_r < 1. \end{cases}$$

In particular, when the counting series is Bernoulli with success probabilities a_j, $j = 1, 2, \ldots, k$, we obtain that $\sum_{j=1}^{k} p_j a_j < 1$.

3.4.5 Random Coefficient INAR(1) model

The random coefficient INAR(1) model is defined in analogy with the existing random coefficients models as

$$X_t = a_{1;t} \circ X_{t-1} + \xi_t,$$

where $\{a_{1;t}\}$ is a stationary process which takes real values. For the case of Bernoulli counting series and $\{a_{1;t}\}$ iid, this class of models has been studied by Zheng et al. (2006) and Zheng et al. (2007). In this case, we can write the above equation as

$$X_t = \sum_{j=1}^{X_{t-1}} Y(U_{t;j}, a_{1;t}) + \varepsilon_t = F(X_{t-1}, \xi_t),$$

where now sequence ξ_t consists of the triplets $(\varepsilon_t, V_t, a_{1;t})$ with $V_t = (U_{t;j})_{j\geq 1}$. Working as before, and using a conditioning argument, we obtain that the conditions for weak dependence are

$$\begin{cases} \|\varepsilon_t\|_r < \infty \\ \|\mathrm{E}(|Y(a_{1;0})|^r \mid \mathcal{F}_{-1})\|_\infty^{1/r} < 1, \end{cases}$$

where the σ-algebra $\mathcal{F}_t = \sigma(\varepsilon_s, V_s, a_{1;s}, s \leq t)$. In particular, when the sequence $\{a_{1,t}\}$ are i.i.d with mean a_1 and the counting series is Bernoulli, then the previous result reduces to the condition $a_1 = \mathrm{E}(|Y(a_{1;0})|) < 1$. The above specification makes evident that a large class of models can be produced in this way; however their dependence conditions are not clear. For instance, long range dependence can be introduced in this way or several other forms of dependence.

3.4.6 Signed Integer-valued Autoregressive (SINAR) models

Following Latour and Truquet (2008) and more recently Kachour and Truquet (2011), define the signed thinning operator by the following. Suppose that $\{Y_i, i \in \mathbb{Z}\}$ is a an i.i.d sequence of integer-valued random variables with cumulative distribution function G. Let X be another integer valued random variable which is independent of Y_i's. Then the signed thinning operator is defined by

$$G \circ X = \begin{cases} \mathrm{sign}(X) \sum_{i=1}^{|X|} Y_i, & \text{if } X \neq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathrm{sign}(x) = 1$, if $x > 0$ and -1 if $x < 0$. This definition generalizes the previous thinning definition and moreover it allows modeling of integer valued time series that assume negative values as well as positive. In particular, (3.8) is generalized by the following signed integer autoregressive process of order p (abbreviated SINAR(p))

$$X_t = \sum_{i=1}^{p} G_i \circ X_{t-i} + \varepsilon_t, \tag{3.9}$$

where the counting sequences $Y_t^{(1)}, \ldots, Y_t^{(p)}$ associated with the c.d.f $G_1, \ldots, G_p$ are mutually independent.

To study the weak dependence properties of process (3.9), it is useful to represent the signed thinning operator in terms of uniform random variables, as in the case of ordinary INAR(p). Towards this goal, suppose that the expectation of the cdf G is a. Then define

$$Y_i = Y(U_i, a) = \sum_{k=0}^{\infty} kI(p_a(k-1) < U_i \leq p_a(k)),$$

by recalling that $p_a(k) = P(Y \leq k)$, $k \in \mathbf{Z}$. Then rewriting (3.9) as

$$\begin{aligned} X_t &= \sum_{i=1}^{p} G_i \circ X_{t-i} + \varepsilon_t = F(X_{t-1}, \ldots, X_{t-p}; \xi_t) \\ &= \sum_{i=1}^{p} \Big\{ \text{sign}(X_{t-i}) \sum_{j=1}^{|X_{t-i}|} Y(U_{t;j} a_i) \Big\} + \varepsilon_t \end{aligned}$$

and applying Theorem 3.2.1, we obtain that the conditions for weak dependence and existence of moments are

$$\begin{cases} \|\varepsilon_t\|_r < \infty \\ \sum_{i=1}^{p} \|Y(G_i)\|_r < 1. \end{cases}$$

When we compare those conditions to the conditions **A1** and **A2** obtained by Kachour and Truquet (2011), we note that they do not restrict the support of the distribution of G_i and ε_t. On the other hand, condition **A2** of Kachour and Truquet (2011) is less strict than the second of the above mentioned conditions.

3.5 Proofs

Proof of Lemma 3.3.1. A crucial step towards our analysis is given by the observation that the indicator function of any point $x_0 \in \mathbb{Z}^d$ can be expressed as

$$1_{x_0}(x) = \begin{cases} 1 - 2d(x, x_0), & \text{if } d(x, x_0) < 1/2 \\ 0, & \text{otherwise,} \end{cases}$$

where $d(\cdot, \cdot)$ is a distance defined in Z^d. However, the function $g(x) = 1 - 2d(x, x_0)$ is a 2-Lipschitz function. From a summation and the fact that all such functions admit disjoint supports, we deduce the useful fact that the same applies

for any discrete set $\mathbb{G} \subset \mathbb{R}^d$. Indeed, let $\mathbb{G}$ be any discrete subset of $\mathbb{R}^d$, $x_0 \in \mathbb{G}$, and consider now the function

$$f_{x_0}(x) = (1 - 2\|x - x_0\|/D)^+, x \in \mathbb{G}.$$

This function is a smooth approximation of the indicator of $x_0 \in \mathbb{G}$ vanishing out of the ball with radius $D/2$. Hence the supports of $f_{x_0}(\cdot)$ and $f_{x_1}(\cdot)$ are disjoint whenever x_0 is not equal to x_1 and both of them belong to $\mathbb{G}$. In other words, we have proved that any function $F : \mathbb{G} \to [-1, 1]$ admits a $\frac{2}{D}-$-Lipschitz extension $\widetilde{F}$ on $\mathbb{R}^d$ defined as

$$\widetilde{F}(x) = \sum_{x_0 \in \mathbb{G}} F(x_0) f_{x_0}(x)$$

□

Proof of Proposition 3.3.1. Suppose that $\{X_t, t \in Z\}$ is and η-dependent process. Then,

$$\begin{aligned} |\mathbb{P}((X_0 \in A) \cap (X_t \in B) - \mathbb{P}(X_0 \in A)\mathbb{P}(X_t \in B)| &= \mathrm{Cov}\,(1_{\{X_0 \in A\}}, 1_{\{X_t \in B\}})| \\ &= |\mathrm{Cov}\,(f_A(X_0), f_B(X_t))| \\ &\leq \frac{4}{D}\eta(t), \end{aligned}$$

where $f_A(\cdot)$ and $f_B(\cdot)$ denote the $\frac{2}{D}$-Lipschitz extensions of indicators of the sets A, B which exist by Lemma 3.3.1. Consider events U, V from the history of the process at negative time which are t-time epochs apart. Suppose that $U = (X_{i_1} \in A_1, \ldots, X_{i_u} \in A_u)$ for times $i_1 \leq i_2 \leq \cdots \leq i_u = 0$ and analogously $V = (X_{j_1} \in B_1, \ldots, X_{j_v} \in B_v)$ for times $t = j_1 \leq i_2 \leq \cdots \leq j_v$. Then the same calculations as before yield

$$|\mathbb{P}(U \cap V) - \mathbb{P}(U)\mathbb{P}(V)| \leq \frac{2}{D}(u + v)\eta(r)$$

□

Proof of theorem 3.3.1. Using Lemma 3.3.1, we obtain for any f and indicator function g, h in $\mathbb{G}^d$, that

$$|\mathbb{E}(f(X_0, X_r) - f(X_0, X_r^*)| = |\mathrm{Cov}\,(g(X_0), h(X_r))| \leq 2\eta(r)/D,$$

similarly to the previous calculations. If now the function f admits a finite support $S \subset \mathbb{G}^{2d}$ then analogously

$$|\mathbb{E}(f(X_0, X_r) - f(X_0, X_r^*)| \leq \frac{2}{D}\mathrm{card}(S)\eta(r),$$

where card$(\cdot)$ denotes cardinality. Finally since the distribution of X_0 is tight, for each $\epsilon > 0$ there exists M_ϵ such that $\mathbb{P}(|X_0| > M_\epsilon) \leq \epsilon$, then replacing f by its restriction to $[-M_\epsilon, M_\epsilon]$ yields

$$|\mathbb{E}(f(X_0, X_r) - f(X_0, X_r^*))| \leq 2\epsilon + 2M_\epsilon^{2d}\eta(r).$$

Therefore, when $\mathbb{E}\|X_0\|^m < \infty$, we derive from Markov inequality that $M_\epsilon \leq (\mathbb{E}\|X_0\|^m/\epsilon)^{1/m}$. One may choose $\epsilon^{1+2/m} = (\mathbb{E}\|X_0\|^m)^{2/m}\eta(r)/D$ to get

$$|\mathbb{E}(f(X_0, X_r) - f(X_0, X_r^*)| \leq 2(\mathbb{E}\|X_0\|^m)^{\frac{2d}{m+2d}} \left(\frac{\eta(r)}{D}\right)^{\frac{m}{m+2d}}$$

We may now consider the case of $p-$Markov processes by setting $\tilde{d} = pd$ and setting $\|(x_1, \ldots, x_{\tilde{d}})\| = \|x_1\| + \cdots + \|x_p\|$ where for $x_1, \ldots, x_p \in \mathbb{G}^d$, and for $u \in \mathbb{G}^d$, $\|u\| = \|(u_1, \ldots, u_d)\| = \max_j |u_j|$. Indeed, $Y_t \equiv (X_t, \ldots, X_{t-p+1}) \in \mathbb{G}^{\tilde{d}}$ is again a Markov chain.

□

Chapter 4

Modeling of DNA Sequence

4.1 Introduction

DNA sequences perform a very important role in the transmission of genetic informations to proteins. Modeling DNA chains is a challenging problem. ACGT stand for the four nucleic acid bases that make up DNA(Adenine, Thymine, Cytosine, Guanine). These four nucleic acids make up a creature's genetic code, or DNA. We aim at classifying and understanding the structure of DNA strings for medical purposes. We consider statistical inference to estimate the distributions of nucleotides under some random hypotheses. In particular, using the strong invariance principle of stochastic processes, this allows to construct SCBs with asymptotically correct nominal coverage probabilities.

We think of the genome as a realization of a stochastic process. A simple model fitting applications is following: we may suppose that the base is A, at the point $t \in [1, n]$ of a DNA string, according to the fact that $U_{t,n} \leq p_A(t/n)$, for $(U_t)_{t\in\mathbb{Z}}$ a process with uniform marginals and where p is the deterministic trend of the model. More generally, functions $p_A, p_C, p_G : [0, 1] \to [0, 1]$ with $0 \leq p_A \leq p_C \leq p_G \leq 1$ provide a model for trends in such strings:

$$\begin{aligned} X_{t,n} = {} & A\mathbb{1}_{\{U_t \leq p_A(\frac{t}{n})\}} + C\mathbb{1}_{\{p_A(\frac{t}{n}) \leq U_t \leq p_A(\frac{t}{n}) + p_C(\frac{t}{n})\}} \\ & + G\mathbb{1}_{\{p_A(\frac{t}{n}) + p_C\frac{t}{n}) \leq U_t \leq p_A(\frac{t}{n}) + p_C(\frac{t}{n}) + p_G(\frac{t}{n})\}} + T\mathbb{1}_{\{U_t > p_A(\frac{t}{n}) + p_C(\frac{t}{n}) + p_G(\frac{t}{n})\}} \end{aligned}$$

Such categorical data rely on specific questions, in order to go back to quantitative data, we suppose $X_{t,n} = \mathbb{1}_{\{U_{t,n} \leq p(t/n)\}}$ as the DNA gene at the point $t \in [1, n]$ is A with $(U_t)_{t\in\mathbb{Z}}$ i.i.d uniform sequence and p is a deterministic trend of the model. Similar models can be proposed for the base C, T and G.

To determine the promoters in DNA strings, one models the fact that at the point $t \in [1, n]$ of DNA the gene is A as $X_{t,n} = \mathbb{1}_{\{U_{t,n} \leq p(t/n)\}}$:

$$X_{t,n} = \left\{ \begin{array}{ll} 1, & \text{with probability } p(\frac{t}{n}) \\ 0, & \text{with probability } 1 - p(\frac{t}{n}) \end{array} \right\}$$

Then $X_{t,n} \sim b(p(t/n))$ follows a Bernoulli distribution with parameter $p(t/n)$.

If there are repeated observations at a fixed point t, Calistri *et al.* (2011) use just the average of the corresponding X_t values to get the estimation of $p(t/n)$. For each genome G they have collected a set of N_G promoter sequences $(X^i_{t,n})_{t\in\mathbb{Z}}$, $i = 1, \ldots, N_G$. A natural idea is to measure the occurrence of A, C, G and T at each position along the aligned set. They are interested in studying the spatial distribution of nucleotides along the promoters by measuring of the percentage of A, C, G and T nucleotides in a set of DNA sequences, i.e.,

$$p_s(t) = \frac{1}{N_G} \sum_{i=1}^{N_G} \mathbb{1}_s(X^i_{t,n})$$

where $s = A, C, G \,\text{and}\, T$.

We would like to provide statistical inference to discuss the asymptotic properties of smoothing methods and the construction of confidence intervals.

We can describe this non stationary time series by a time-varying model,

$$X_{t,n} = p(\frac{t}{n}) + \sqrt{p(\frac{t}{n})(1 - p(\frac{t}{n}))}\xi_t, \;\; t = 1, \ldots, n. \tag{4.1}$$

where ξ_t admits the mean 0 and the variance 1. ξ_t is non i.i.d but it is a weak white noise in $\mathbb{L}^2$ (i.e $\mathbb{E}\xi_t = 0$ and $\mathbb{E}\xi_t^2 = 1$). The support of ξ_t also depends on $t \in [0, 1]$. For the sake of simplicity, we denote

$$X_{t,n} = p(\frac{t}{n}) + \sigma(\frac{t}{n})\xi_t \quad \text{with} \quad \sigma^2(t) = p(\frac{t}{n})(1 - p(\frac{t}{n})).$$

The process X_t is non-stationary and can be interpreted as a signal plus noise model. The objective is to describe this sequence by modeling the process and

testing the proposed model. Since the mean of X_t varies over time, we estimate this trend in a first time. Interesting special features are, for instance, monotonicity or convexity.

Those trends are determinant for individuals. Standard kernel-type smoothing techniques are processed together with the development of asymptotic in this case. Asymptotic properties of nonparametric estimates for time series have been widely discussed under various strong mixing conditions; see Robinson (1983), Bosq (1996), Doukhan and Louhichi (1999) among others.

In this chapter, we provide central limit theorems for the kernel-type estimator to the case of general process satisfying the strong invariance principle conditions. For our goal of constructing SCBs for p, we assume that p is smooth. SCBs can be used to find parametric forms of p. For example, in the study of global temperature series, an interesting problem is to test whether the trend is linear, quadratic or of other patterns. Applying the strong invariance principle of stochastic processes, we shall provide a solution to the problem and construct SCBs with asymptotically correct nominal coverage probabilities. Another interesting problem is to test the monotone or convexity of the trends p. We point out that if it were possible to model genome sequences as stochastic process, one could construct a test for monotone or convexity based on the asymptotically correct nominal coverage for p' and p''.

Our starting point is the same as in Wu and Shao (2007). Those authors prove that a strong approximation principle for the partial sums of a stationary process with an explicit rate to entails simultaneous confidence bands with asymptotically correct nominal coverage probabilities. In their paper, they point out that an explicit rate in the strong approximation principle is crucial to control certain errors terms (see their Remark 2). The possible bandwidth heavily depend on the previous convergence rate.

4.2 Mains results

4.2.1 Asymptotic properties

We begin in this section by introducing our estimators. Let K be a real-valued, bounded and kernel function with $\int K(u)\mathrm{d}u = 1$. There exists a vast literature on

nonparametric estimation of the regression function p. Here we use the Priestley-Chao estimator

$$\hat{p}_{h_n}(t) = \sum_{i=1}^{n} \frac{1}{nh_n} K\left(\frac{t - i/n}{h_n}\right) X_{i,n}. \tag{4.2}$$

The bandwidth $h_n \to 0$ satisfies $nh_n \to \infty$. Some regularity conditions on K are imposed below.

Our object will be to get global measures of how good $\hat{p}_{h_n}(t)$ is as an estimate of $p(t)$. We assume $\int K(u)u\mathrm{d}u = 0$ and $\int K(u)u^2\mathrm{d}u \neq 0$. $p(t)$ must be twice differentiable. Then it is known that if $h_n \to 0$ as $n \to \infty$ in such a way that $nh_n \to \infty$,

$$E(\hat{p}_{h_n}(t)) \sim p(t) + \frac{1}{2}h_n^2 d_K p''(t) \tag{4.3}$$

and if $\sum_{i=-\infty}^{\infty} \mathbb{E}\xi_0\xi_i \leq \infty$, one obtains

$$\mathrm{Var}\,(\hat{p}_{h_n}(t)) \sim \frac{\sigma^2(t)\gamma c_K}{nh_n} \tag{4.4}$$

where $c_K = \int K(u)^2\mathrm{d}u$, $d_K = \int_{-1}^{1} K(u)u^2\mathrm{d}u$ and $\gamma = \sum_{i=-\infty}^{\infty} \mathbb{E}\xi_0\xi_i$.

Quite often the regression curve itself is not the target of interest but rather derivatives of it. The technique of kernel estimation can also be used to estimate derivatives of the regression function. Kernel derivative estimators are defined by differentiating the kernel function with respect to t. If the kernel is sufficiently smooth and the bandwidth sequence is correctly tuned then these estimators will converge to the corresponding derivatives.

Definition 4.2.1. *Let the function p be a $\mathcal{C}^q[0,1]$ function, with bounded derivations, for some $q \in \mathbb{N}^*$. Then the k-th($k \leq q$) derivative with respect to t gives*

$$\hat{p}_{h_n}^{(k)}(t) = n^{-1}h_n^{-(k+1)} \sum_{i=1}^{n} K^{(k)}\left(\frac{t - i/n}{h_n}\right) X_{i,n}$$

The derivative estimator $\hat{p}_{h_n}^{(k)}(t)$ is asymptotically unbiased. Assume that for some $q \in \mathbb{N}^*$, the function K be a $\mathcal{C}^q[0,1]$ function with $K^{(j)}(0) = K^{(j)}(1) = 0$, $j = 0, \ldots, q-1$. Then elementary calculations show that

$$E(\hat{p}_{h_n}^{(k)}(t)) \sim p^{(k)}(t) + h_n^2 d_K^{(k)} p^{(k+2)}(t)/(k+2)!$$

The variance of $E(\hat{p}_{h_n}^{(k)}(t))$ tends to zero if $nh_n^{2k+1} \to \infty$,

$$\mathrm{Var}\,(\hat{p}_{h_n}^{(k)}(t)) \sim \frac{\sigma^2(t)\gamma c_K^{(k)}}{nh_n^{2k+1}}$$

where $d_K^{(k)} = \int K^{(k)}(u)u^{k+2}\mathrm{d}u$ and $c_K^{(k)} = \int K^{(k)}(u)^2\mathrm{d}u$.

If K and its derivatives are Lipschitz continuous and have bounded support, elementary calculations show that Theorem (4.2.1) (4.2.2) (4.2.3) assert central limit theorems (CLT) for $\hat{p}_{h_n}(t)$, $\hat{p'}_{h_n}(t)$ and $\hat{p''}_{h_n}(t)$, which can be used to construct point-wise confidence intervals for $p(t)$ $p'(t)$ and $p''(t)$.

Assumption SIP: Let $(\xi_i)_{i\in\mathbb{Z}}$ be some centered dependent process with a finite second moment, there exists a sequence $(Z_i)_{i\geq 1}$ of i.i.d centered Gaussian variables such that

$$\sup_{i\leq k\leq n} |\sum_{i=1}^{k}(\xi_i - Z_i)| = o_{\mathrm{AS}}(n^{\alpha}\log n) \quad 1/4 \leq \alpha \leq 1/2 \tag{4.5}$$

.

Example 4.2.1 (Causal Bernoulli shifts)**.** *Let $(\xi_n)_{n\in\mathbb{Z}}$ defined by*

$$\xi_n = H(\varepsilon_n, \varepsilon_{n-1}, \varepsilon_{n-2}, \dots,)$$

where $\varepsilon_i, i \in \mathbb{Z}$ are iid random variables and H is a measurable function such that ξ_i is well-defined. By interpreting causal Bernoulli shifts as physical systems, Wu (2005) introduces physical dependence coefficients quantifying the dependence of outputs (ξ_t) on inputs (ε_t). Let ε'_j be an IID copy of ε_j and $\xi_n^* = H(\varepsilon_n, \varepsilon_{n-1}, \dots, \varepsilon_1, \varepsilon'_0, \varepsilon_{-1}, \dots)$. Assume that $\mathbb{E}\|\xi_n\|_m < \infty$, $m > 2$, he considers the nonlinear system theory's coefficient

$$\delta_m(n) = \|\xi_n - \xi_n^*\|_m.$$

For a variety of non-linear time series models, There exists $r \in (0,1)$ such that

$$\delta_m(n) = \|\xi_n - \xi_n^*\|_m = \mathcal{O}(r^n)$$

Wu (2007) showed that under $\sum_{i=1}^{\infty} i\|\xi_n - \xi_n^*\|_m < \infty$, the condition 4.5 holds.

Example 4.2.2 (Symmetric random walk on the circle)**.** *Let us define the Markov kernel K_M by $K_M f(x) = \frac{1}{2}(f(x+a) + f(x-a))$ on the torus $\mathbb{R}/\mathbb{Z}$, with a irrational in $[0,1]$, and the Lebesgue-Haar measure μ is the unique probability which is invariant by K_M. We assume that $(\varepsilon_i)_{i\in\mathbb{Z}}$ is the stationary Markov chain with transition kernel K_M and invariant distribution μ. For $f \in \mathbb{L}^2(\mu)$, let*

$$\xi_k = f(\varepsilon_k) - \mu(f).$$

Let a satisfy $\min_{i\in\mathbb{Z}} |ka - i| \geq c(a)|k| - 1$ for some positive constant $c(a)$ and $\hat{f}(k)$ be the Fourier coefficients of f. Assume that for some positive ϵ,

$$\sup_{k\neq 0} |k|^s \left(\log(1+|k|)\right)^{1+\epsilon} |\hat{f}(k)| < \infty \text{ where } s = \sqrt{\alpha^2 - 2\alpha + 4} - 3\alpha + 2.$$

then the condition 4.5 holds with $\sigma^2 = \sum_k Cov(\xi_0, \xi_k)$.(see Dedecker et al. (2012))

Assumption $\mathcal{H}(a)$: Let $\mathcal{H}(a), 1 \leq a \leq 2$, be the set of bounded functions H with bounded support satisfying

1. $\int_{\mathbb{R}} \Psi_H(u,\delta)\mathrm{d}u = \mathcal{O}(\delta)$ as $\delta \to 0$, where $\Psi_H(u,\delta) = \sup\{|H(y) - H(y')| : y, y' \in [u-\delta, u+\delta]\}$, and,
2. the $\lim D_{H,a} = \lim_{\delta\to 0}[|\delta|]^{-a} \int_{\mathbb{R}}\{H(x+\delta) - H(x)\}^2\mathrm{d}x$ exists and $D_{H,a} \neq 0$.

For $m \geq 3$ define

$$B_{H,a}(m) = \sqrt{2\log(m)} + \frac{1}{\sqrt{2\log m}}\left[\frac{2-a}{2a}\log\log m + \log\left(\frac{C_{H,a}^{1/a} h_a 2^{1/a}}{2\sqrt{\pi}}\right)\right].$$

where $C_{H,a} = D_{H,a}/2\int_{\mathbb{R}} H^2(s)\mathrm{d}s$ and h_a has two values $h_1 = 1$ and $h_2 = \pi^{-1/2}$ (see Bickel and Rosemblatt (1973),Wu and Shao (2007)).

Example 4.2.3. *The triangle, quartic, Epanechnikov and Parzen kernels satisfies the previous assumptions with $a = 2$, and $a = 1$ for the rectangle kernel.*

Theorem 4.2.1. *Let us assume that* **Assumption SIP** *is satisfied and that K has bounded variation, $h_n \to 0$ and $(\log n)^2 = o(n^{1-2\alpha}h_n)$. Then for fixed $0 < t < 1$,*

$$\sqrt{nh_n}\left\{\hat{p}_{h_n}(t) - \mathbb{E}\hat{p}_{h_n}(t)\right\} \xrightarrow{\mathbb{P}} \mathcal{N}(0, \sigma^2(t)\gamma c_K).$$

Now, following the regularity of the function $p, \mathbb{E}\hat{p}(t)$ is a more or less good approximation of $p(t)$. Hence, here we provide an approximation of the bias. Let $\mathcal{C}^q([0,1])$, $q = 0, 1, \ldots, ,$ denote the collection of functions having un to q-th order derivatives.

Corollary 4.2.1. *Assume that for some $q \in \mathbb{N}^*$, the function p is a $\mathcal{C}^q[0,1]$ function, with bounded derivations. Then, under the conditions of Theorem*

(4.2.1) , *with* K *a kernel such that* $\int K(u)u^s \mathrm{d}u = 0$ *for* $s = \{1, \ldots, q-1\}$ *and* $\int K(u)u^q \mathrm{d}u \neq 0$, *if* $h_n = C \cdot n^{-1/(2q+1)}$ *(with* $C > 0$*) then*

$$\sqrt{nh_n}\left\{\hat{p}_{h_n}(t) - p(t)\right\} \xrightarrow{\mathbb{P}} \mathcal{N}(p^{(q)}(t)\frac{1}{q!}\int u^q K(u)\mathrm{d}u, \sigma^2(t)\gamma c_K).$$

Replacing $\hat{p}(t)$ with $\hat{p^{'}}_{h_n}(t)$, $\hat{p^{''}}_{h_n}(t)$ in the prove of 4.2.1 lead us Theorem 4.2.2 and 4.2.3 below.

Theorem 4.2.2. *Assume that* **Assumption SIP** *is satisfied, Let* K *be a function in* $\mathcal{C}^1[0,1]$ *and* K' *has bounded variation,* $h_n \to 0$, $nh_n^3 \to \infty$ *and* $(\log n)^2 = o(n^{1-2\alpha}h_n)$. *Then for fixed* $0 < t < 1$,

$$\sqrt{nh_n^3}\left\{\hat{p}'_{h_n}(t) - \mathbb{E}\hat{p}'_{h_n}(t)\right\} \xrightarrow{\mathbb{P}} \mathcal{N}(0, \sigma^2(t)\gamma c_K^{(1)}).$$

Theorem 4.2.3. *Assume that* **Assumption SIP** *is satisfied, Let* K *be a function in* $\mathcal{C}^2[0,1]$ *and* K'' *has bounded variation ,* $h_n \to 0$, $nh_n^5 \to \infty$ *and* $(\log n)^2 = o(n^{1-2\alpha}h_n)$. *Then for fixed* $0 < t < 1$,

$$\sqrt{nh_n^5}\left\{\hat{p''}_{h_n}(t) - \mathbb{E}\hat{p''}_{h_n}(t)\right\} \xrightarrow{\mathbb{P}} \mathcal{N}(0, \sigma^2(t)\gamma c_K^{(2)}).$$

To construct an asymptotic SCB for $p(t)$ over the interval $t \in \mathcal{T}$ with level $(1-\alpha)$, $\alpha \in (0,1)$, We need to find two functions $l_n(t)$ and $u_n(t)$ based on the data such that

$$\lim_{n\to\infty} P(l_n(t) \leq p(t) \leq u_n(t), \text{ for all } t \in \mathcal{T}) = 1 - \alpha.$$

A closely related problem is to study the asymptotic uniform distributional theory for the estimator $\hat{p}_{h_n}(t)$. Namely, one needs to find the asymptotic distribution for $\sup_{0<t<1} |\hat{p}_{h_n}(t) - \mathbb{E}\hat{p}_{h_n}(t)|$.

Theorem (4.2.4) (4.2.5) (4.2.6) provide theoretical SCBs for p p' and p'' with asymptotically correct coverage probabilities under slightly different model.

The construction of SCB l_n and u_n has been a difficult problem if dependence is present. A key tool in Wu's approach is Bickel and Rosemblatt (1973) asymptotic theory for maximal deviations of kernel density estimators Bickel and Rosenblatt applied a deep result in probability theory, strong approximation, which asserts that normalized empirical processes of independent random variables can be approximated by Brownian bridges. Mention that both Wu (2007) and Dedecker *et*

al. (2012) give rates of convergence in the strong invariance principle for stationary sequences satisfying some projective criteria. Thus one can construct simultaneous confidence bands with asymptotically correct nominal coverage probabilities for time series.

Theorem 4.2.4 (Wu and Shao (2007))**.** *Assume that* **Assumption** $\mathcal{H}(a)$ *is satisfied for* K *and that* K *is a symmetric kernel with support* $[-\omega, \omega]$*. Further assume that* $p \in \mathcal{C}^3[0,1]$ *and*

$$\frac{(\log(n))^3}{h_n n^{1-2\alpha}} + nh_n^7 \log(n) \to 0. \tag{4.6}$$

Let $m = 1/h_n$ *and the interval* $\mathcal{T} = [\omega h_n, 1 - \omega h_n]$*. Then for every* $u \in \mathbb{R}$*, as* $n \to \infty$*,*

$$\mathbb{P}\left[\frac{\sqrt{nh_n}}{\gamma c_K \sqrt{c_K}} \sup_{t\in\mathcal{T}} \frac{1}{\sigma(t)} |\hat{p}_{h_n}(t) - p(t) - \frac{1}{2}h_n^2 p''(t) d_k| - B_{K,a}(m) \le \frac{u}{\sqrt{2\log m}}\right] \to \exp\{-2\exp(-u)\}$$

In condition (4.6), the first part ensures the validity of the strong approximation and the second part controls the bias. Condition (4.6) are satisfied if $h_n \asymp n^{-\gamma}$, $1/7 < \gamma < 1/2$. Indeed, the choir of $\gamma = 1/5$ is well known as the optimal bandwidth under the mean-squared error criterion. Analogue results may be provided for p' and p''.

Theorem 4.2.5. *Assume that* **Assumption** $\mathcal{H}(a)$ *is satisfied for* K' *and that* K' *is a symmetric kernel with support* $[-\omega, \omega]$*. Further assume that* $p \in \mathcal{C}^4[0,1]$ *and*

$$\frac{(\log(n))^3}{h_n n^{1-2\alpha}} + nh_n^9 \log(n) \to 0. \tag{4.7}$$

Let $m = 1/h_n$ *and the interval* $\mathcal{T} = [\omega h_n, 1 - \omega h_n]$*. Then for every* $u \in \mathbb{R}$*, as* $n \to \infty$*,*

$$\mathbb{P}\left[\frac{\sqrt{nh_n^3}}{\gamma c_K^{(1)} \sqrt{c_K^{(1)}}} \sup_{t\in\mathcal{T}} \frac{1}{\sigma(t)} |\hat{p}'_{h_n}(t) - p'(t) - \frac{1}{6}h_n^2 p^{(3)}(t) d_k^{(1)}| - B_{K',a}(m) \le \frac{u}{\sqrt{2\log m}}\right] \to \exp\{-2\exp(-u)\}$$

We also have $nh_n^3 \to \infty$ for the estimator $\hat{p}'_{h_n}(t)$, combine the condition 4.6, $h_n \asymp n^{-\gamma}$ with $1/7 < \gamma < 1/3$.

Theorem 4.2.6. *Assume that* **Assumption** $\mathcal{H}(a)$ *is satisfied for* K'' *and that* K'' *is a symmetric kernel with support* $[-\omega, \omega]$. *Further assume that* $p \in \mathcal{C}^5[0,1]$ *and*

$$\frac{(\log(n))^3}{h_n n^{1-2\alpha}} + nh_n^{11}\log(n) \to 0. \tag{4.8}$$

Let $m = 1/h_n$ *and the interval* $\mathcal{T} = [\omega h_n, 1 - \omega h_n]$. *Then for every* $u \in \mathbb{R}$, *as* $n \to \infty$,

$$\mathbb{P}\left[\frac{\sqrt{nh_n^5}}{\gamma c_K^{(2)}\sqrt{c_K^{(2)}}}\sup_{t\in\mathcal{T}}\frac{1}{\sigma(t)}|\hat{p''}_{h_n}(t) - p''(t) - \frac{1}{24}h_n^2 p^{(4)}(t)d_k^{(2)}| - B_{K'',a}(m) \le \frac{u}{\sqrt{2\log m}}\right]$$
$$\to \exp\{-2\exp(-u)\}$$

Note that $h_n \asymp n^{-\gamma}$ with $1/7 < \gamma < 1/5$, since $nh_n^5 \to \infty$.

4.2.2 Hypothesis Testing

Calistri *et al.* (2011) noticed that the frequencies of A, C, G and T nucleotides in a set of DNA sequences on bacteria appear monotone or convexity. So we now focus on developing tests of statistical significance for the monotone and convexity property of the genome sequences. The test of monotone or convexity is related to first or second order derivatives. We shall see that this hypothesis is not satisfied in any of the cases we studied.

We first assume that the function to test, under both hypotheses, belongs to a certain class of regular function. The hypothesis of positivity is referred to as null hypothesis. this hypothesis is composite and presented by

$$H_0 : \inf_{0\le t\le 1} g(t) \ge 0$$

and the alternative hypothesis is defined as:

$$H_1 : \inf_{0\le t\le 1} g(t) < 0.$$

The test is set up so that H_0 is rejected (H_1 is accepted) at significance level α if $\mathbb{P}_{H_0}(\inf_t \hat{g}(t) \le \epsilon(n)) \le \alpha$ where $\epsilon(n)$ is chosen such that the probability of a type I error is less than or equal to α. For some $\epsilon(n) > 0$, the probability of a type I error may be written as

$$\mathbb{P}(H_0 \text{ is rejected}|H_0 \text{ is true}) = \mathbb{P}(\inf_{0<t<1}\hat{g}(t) \le -\epsilon(n)|\inf_{0<t<1} g(t) \ge 0)$$

$$\leq \mathbb{P}\left(\sup_t |\hat{g}(t) - g(t)| \geq \epsilon(n)\right) \to \alpha$$

As we know that $\hat{g}(t)$ converge to $g(t)$ in probability, the power of the test is written for $g \in H_1$,

$$\mathbb{P}(\inf_{0<t<1} \hat{g}(t) \leq -\epsilon(n)) \to 1$$

Proposition 4.2.1 and 4.2.2 are simple consequences of Theorem 4.2.5 and 4.2.6.

Proposition 4.2.1. *Under the assumptions of Theorem 4.2.5, choose* $0 < \alpha < 1$, *let* $\epsilon(n) = \left(\frac{-\log(-\frac{1}{2}\log(1-\alpha))}{\sqrt{2\log m}} + B_{K',a}(m)\right) \frac{\gamma c_K^{(1)} \sqrt{c_K^{(1)}}}{\sqrt{nh_n^3}} \sup_t \sigma(t)$, *then we have for* $n \to \infty$

$$\mathbb{P}\left(\sup_t |\hat{p'}_{h_n}(t) - p'(t)| \geq \epsilon(n)\right) \leq \alpha$$

and

$$\lim_{n\to\infty} \mathbb{P}(\inf_{0<t<1} \hat{p'}_{h_n}(t) \leq -\epsilon(n)) \to 1$$

Proposition 4.2.2. *Under the assumptions of Theorem 4.2.6, choose* $0 < \alpha < 1$, *let* $\epsilon(n) = \left(\frac{-\log(-\frac{1}{2}\log(1-\alpha))}{\sqrt{2\log m}} + B_{K'',a}(m)\right) \frac{\gamma c_K^{(2)} \sqrt{c_K^{(2)}}}{\sqrt{nh_n^5}} \sup_t \sigma(t)$, *then we have for* $n \to \infty$

$$\mathbb{P}\left(\sup_t |\hat{p''}_{h_n}(t) - p''(t)| \geq \epsilon(n)\right) \leq \alpha$$

and

$$\lim_{n\to\infty} \mathbb{P}(\inf_{0<t<1} \hat{p''}_{h_n}(t)) \leq -\epsilon(n)) \to 1$$

4.2.3 Implementation

Let us detail our specific proposal for confidence band. Let $\hat{\sigma}_{b_n}(t)$ be estimates of σ. On the basis of theorem (4.2.4), The asymptotic $100(1-\alpha)\%$ confidence band for p we use take the form

$$\hat{p}_{h_n}(t) - h_n^2 \beta \hat{p}''(t) \pm \ell_{u_\alpha},$$

where

$$\ell_{u_\alpha} = \frac{\hat{\sigma}_{h_n}(t)\gamma c_K}{\sqrt{(nh_n)}} \left[B_{K,a}(h_n^{-1}) + \frac{-\log\left(\log(1-\alpha)^{-1/2}\right)}{\sqrt{(2\log(h_n^{-1}))}}\right]$$

To construct the confidence band it requires the knowledge of $\hat{p}''$, which cannot be easily estimated. Following Wu and Shao (2007) , we adopt a jackknife-type bias correction scheme which avoids estimating $\hat{p}''$:

$$\hat{p}^*_{h_n}(t) = 2\hat{p}_{h_n}(t) - \hat{p}_{\sqrt{2}h_n}(t)$$

This is equivalent to using the higher (4-th) order kernel

$$K^*(u) = 2K(u) - \frac{K(u/\sqrt{2})}{\sqrt{2}}$$

The bias term $O(h_n^2)$ in $\hat{p}_{h_n}(t)$ reduces to $O(h_n^4)$ in $\hat{p}^*_{h_n}(t)$. Remark that this is particularly convenient as we only estimate the $\hat{p}^*_{h_n}(t)$ once and we can use it to approximate the confidence bands.

Definition 4.2.2. *Recall that $X_t \sim b(p(t/n))$ follows a Bernoulli distribution with parameter $p(t/n)$, so*

$$\hat{\sigma}^2(t) = \hat{p}_{h_n}(t)(1 - \hat{p}_{h_n}(t))$$

Replacing $\sigma^2(t)$ with $\hat{\sigma}^2_{h_n}(t)$ gives the approximate confidence intervals that is applicable in practice.

It is well known that the convergence to the extreme value distributions in 4.2.4 is extremely slow and very large values of n are needed for the approximation to be reasonably accurate. We shall propose a finite sample approximation scheme to compute the cutoff value q_α. Let Z_i, $1 \leq i \leq n$, be i.i.d. standard normal random variables, model (4.1) can be reduced to the convention model

$$\bar{X}_{k,n} = p(\frac{t}{n}) + \sigma(k)Z_k, \ k = 1, \ldots, n.$$

So we propose the finite sample cutoff value q_α defined by

$$\mathbb{P}\{\sup_{1\leq i\leq n} |Z_i| < q_\alpha\} = 1 - \alpha.$$

4.3 Simulation study

In this section, a simulation study shall be given for the performance of our estimators and SCBs in section 2.2.3. We choose the mean function $p(t) = \sin(\frac{\pi}{2}t)$ with $t = 1, \ldots, n$, and consider the model

$$X_{t,n} = \left\{ \begin{array}{ll} 1, & \text{with probability } p(\frac{t}{n}) \\ 0, & \text{with probability } 1 - p(\frac{t}{n}) \end{array} \right\}$$

Let $n = 1000$, to estimate $q_{0.95} = q_{0.95}(h)$ for each b, we draw an iid sample $Z_1, \ldots, Z_n$ from the normal standard distribution, and calculate $\sup_{0\le t\le 1} |\bar{p}_h^*(t)|$, where $\bar{p}_h^*(t) = 2\bar{p}_h(t) - \bar{p}_{h\sqrt{2}}(t)$ and $\bar{p}_h(t) = \sum_{i=1}^n \frac{1}{nh} K\left(\frac{t-i/n}{h}\right) Z_i$. The estimated quantile $\hat{q}_{0.95}$ is obtained by generating $N = 10^4$ realization of $\bar{p}_h^*(t)$. The 95% SCB is constructed as $\hat{p}*_h(t) \pm \hat{\sigma}_{h_n}(t)\hat{q}_{0.95}$. For $\alpha = 0.05$, $q_{0.95} = 0.308$ and the optimal bandwidths is $h_n = 0.20$, choosing by the kernel regression smoothing program *glkerns* in the R package. Figure 4.1 and Figure 4.2 report the results.

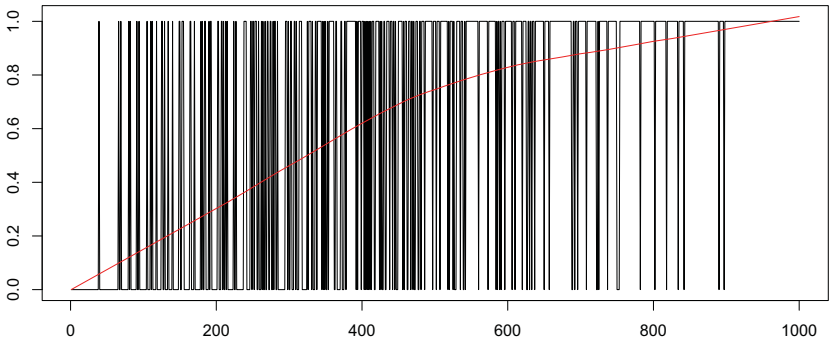

Figure 4.1: an kernel estimator for p(t).

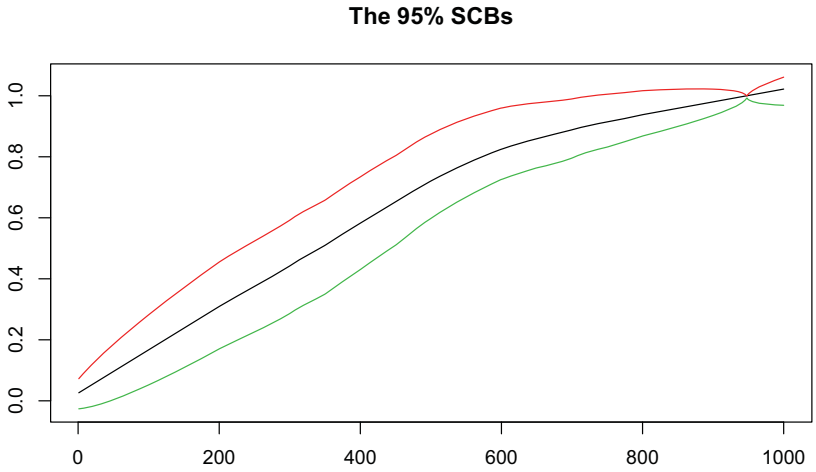

Figure 4.2: an asymptotic SCB for p(t) with level 95%.

To test the monotone, we choose $a = 1$, $c_K^{(1)} = c_K^{(2)} = 1/2$, then $\epsilon(n) = 0.077$. The test $H_0 : \inf_t p'(t) \geq 0$ is accepted with $\inf_t \hat{p}'_{h_n}(t) = 0.081 > -\epsilon(n)$.

4.4 Application

Here we consider the series $(X_i)_{1 \leq i \leq 1000}$ of nucleotide of an eucaryote. The purpose is to estimate the trends and give an asymptotic SCB . We shall use the simulation method in 4.3 to obtain cut-off values. Let $n = 1000$. We repeat the following process for 10^4 times: generate n iid normals $\mathcal{N}(0, 1)$ and calculate $\bar{p}_h^*(t)$. The 95% and 99% simulated quantiles are 0.39 and 0.42 respectively.

Figure 4.3 and 4.4 show our asymptotic SCB for the trends of DNA data with level 95% and 99%. The test statistic $\inf_t \hat{p}'_{h_n}(t) = -1.005 <= -\epsilon(n) = -0.028$, and $\inf_t \hat{p}''_{h_n}(t) = -17.153 <= -\epsilon(n) = -0.073$.

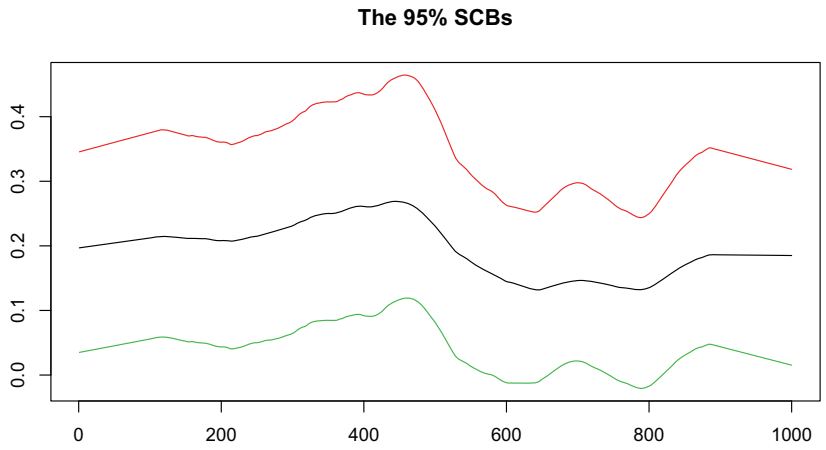

Figure 4.3: an asymptotic SCB for the trends of DNA data with level 95%.

As we have seen so far, in eukaryotes, while remaining constant in the upstream part of the analyzed regions. the trend of the nucleotide base A changes downstream.

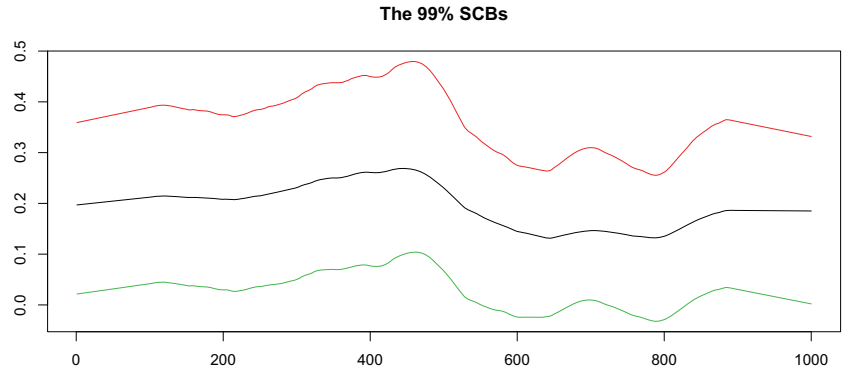

Figure 4.4: an asymptotic SCB for the trends of DNA data with level 99%.

4.5 Some Preliminary Lemmas and Proofs

Lemma 4.5.1 (Wu and Shao (2007)). *Assume that* $H \in \mathcal{H}(a)$, $a \in [1,2]$, $\int_{\mathbb{R}} H^2(u) = 1$ *and* H *has finite support* $[-\omega, \omega]$. *Let* $h_n \to 0$ *satisfy* $\sqrt{nh_n}/(\log n)^3 \to \infty$. *For* $0 \le t \le 1$ *define*

$$U_n(t) = \frac{1}{\sqrt{nh_n}} \sum_{j=1}^{n} H\Big(m(t - \frac{j}{n})\Big) \frac{e_j}{\sigma(t)}$$

where $m = 1/h_n$. *Then, for* $u \in \mathbb{R}$

$$\lim_{n\to\infty} (\mathbb{P}[\max_{t\in[\omega h_n, 1-\omega h_n]}] - B_{H,a}(m) \le \frac{u}{\sqrt{2\log m}}) = \exp(-2\exp(-u))$$

Lemma 4.5.2 (Wu and Shao (2007)). *Let* $K \in \mathcal{H}(a)$ *be a symmetric kernel with support* $[-\omega, \omega]$ *and* $p \in \mathcal{C}^3[0,1]$. *Then* $\mathbb{E}(p_{h_n}(t)) - p(t) = \frac{1}{2}h_n^2 p''(t) d_K + O(h_n^3 + (nh_n)^{-1})$ *uniformly over* $t \in \mathcal{T} = [\omega h_n, 1 - \omega h_n]$.

Lemma 4.5.3. *Let* K *be a* $\mathcal{C}^q[0,1]$ *symmetric kernel function with support* $[-\omega, \omega]$ *for some* $q \in \mathbb{N}^*$. *Let* $K^{(k)} \in \mathcal{H}(a)$ *and* $p \in \mathcal{C}^q[0,1]$. *Then for* $k \le q$ $\mathbb{E}(p_{h_n}^{(k)}(t)) - p^{(k)}(t) = \frac{h_n^2}{(k+2)!} p^{(k+2)}(t) d_K^{(k)} + O(h_n^3 + (nh_n)^{-1})$ *uniformly over* $t \in \mathcal{T} = [\omega h_n, 1 - \omega h_n]$.

The proof is similar to Lemma 4.5.2 and the details are omitted.

Proof of Theorem (4.2.1). Let $(\xi_i)_{i\in\mathbb{Z}}$ be a dependent time series with real values, zero mean and variance 1. Assume that $\|\xi_0\|_\infty < \infty$ and $\gamma = \sum_{i=-\infty}^{\infty} \mathbb{E}(\xi_0 \xi_i) <$

∞. Denote

$$\hat{p}_{h_n}(t) = \sum_{j=1}^{n} \omega_n(t,j) X_{j,n}$$

where $\omega_n(t,j) = \frac{1}{nh_n} K\left(\frac{t-j/n}{h_n}\right)$ are suitable weights.

Let $t \in [0,1]$, we have

$$Y(t) = \hat{p}_{h_n}(t) - \mathbb{E}\hat{p}_{h_n}(t) = \sum_{j=1}^{n} \omega_n(t,j)\sigma(\frac{j}{n})\xi_j$$

we now define the Gaussian process

$$Y^*(t) = \sum_{j=1}^{n} \omega_n(t,j)\sigma(\frac{j}{n})Z_j$$

using the summation by parts formula, we have

$$|Y(t) - Y^*(t)| \leq \Omega(t) \sup_{k \leq n} |\sum_{i=1}^{k} (\xi_i - Z_i)| = o_{\mathrm{AS}}(\Omega(t) n^{\alpha} \log n)$$

where $\Omega(t) = |\omega_n(t,1)\sigma(\frac{1}{n})| + \sum_{k=1}^{n-1} |\left(\omega_n(t,k+1)\sigma(\frac{k+1}{n}) - \omega_n(t,k)\sigma(\frac{k}{n})\right)|$.

Let $\Omega_n = \max_{0 \leq t \leq t} \Omega(t)$, we obtain the uniform approximation

$$\|Y(t) - Y^*(t)\|_{\infty} = o_{\mathrm{AS}}(\Omega_n n^{\alpha} \log n)$$

If K has bounded variation $\Omega_n(t)$ have tractable bounds and $\Omega_n = O\left((nh_n)^{-1}\right)$. Thus with $(\log n)^2 = o(n^{1-2\alpha} h_n)$,

$$\sqrt{nh_n} \|Y(t) - Y^*(t)\|_{\infty} \xrightarrow{\mathbb{P}} 0$$

□

Proof of Corollary (4.2.1). Under the assumption on K and p is a $\mathcal{C}^q(\mathbb{R})$ function for some $q \in \mathbb{N}^*$,

$$\mathbb{E}(\hat{p}_{h_n}(t)) = p(t) + h_n^q \cdot (1 + o(1)) \cdot p^{(q)}(t) \frac{1}{q!} \int u^q K(u) \mathrm{d}u$$

It implies the optimal choice convergence rate of h_n. □

Proof of Theorem (4.2.4). By condition 4.6, $(h_n^3 + (nh_n)^{-1})\sqrt{nh_n} = o(\sqrt{\log n})$, and the Theorem follows from Lemma 4.5.1 and 4.5.2, which concern the stochastic part $p_{h_n}(t) - \mathbb{E}(p_{h_n}(t))$ and the bias $\mathbb{E}(p_{h_n}(t)) - p(t) = \frac{2}{1} h_n^2 p''(t) d_K + O(h_n^3 + (nh_n)^{-1})$ respectively. □

Proof of Theorem (4.2.5). By condition 4.7, $(h_n^3 + (nh_n)^{-1})\sqrt{nh_n^3} = o(\sqrt{\log n})$, and the Theorem follows from Lemma 4.5.1 and 4.5.3. □

Proof of Theorem (4.2.6). By condition 4.8, $(h_n^3 + (nh_n)^{-1})\sqrt{nh_n^5} = o(\sqrt{\log n})$, and the Theorem follows from Lemma 4.5.1 and 4.5.3. □

Part II

Time Series Forecasting under Weak Dependence Conditions

Chapter 5

Prediction of Time Series by Statistical Learning

The aim of this part is the study of statistical properties of learning algorithm in the case of time series prediction. A series of papers (e.g. Meir (2000); Modha and Masry (1998); Alquier and Wintenberger (2012)) extends the oracle inequalities obtained for i.i.d observations to time series under weak dependence conditions. Given a family of predictors and n observations, oracle inequalities state that a predictor forecasts the series as well as the best predictor in the family up to a remainder term Δ_n. Using the PAC-Bayesian approach, we establish under weak dependence conditions oracle inequalities with optimal rates of convergence Δ_n for Gibbs estimator. Similar results were proved for the ERM procedure under a restriction on the parameter space. We apply the method for quantile forecasting of the french GDP with promising results.

5.1 Introduction

Motivated by economics problems, the prediction of time series is one of the most emblematic problems of statistics. Various methodologies are used that come from such various fields as parametric statistics, statistical learning, computer science or game theory.

In the parametric approach, one assumes that the time series is generated from a parametric model, e.g. ARMA or ARIMA, see Hamilton (1994); Brockwell and Davis (2009). It is then possible to estimate the parameters of the model

and to build confidence intervals on the prevision. However, such an assumption is unrealistic in most applications.

In the statistical learning point of view, one usually tries to avoid such restrictive parametric assumptions - see, e.g., Cesa-Bianchi and Lugosi (2006); Stoltz (2009) for the online approach dedicated to the prediction of individual sequences, and Modha and Masry (1998); Meir (2000); Alquier and Wintenberger (2012) for the batch approach. However, in this setting, a few attention has been paid to the construction of confidence intervals or to any quantification of the precision of the prediction. This is a major drawback in many applications. Notice however that Biau and Patra (2011) proposed to minimize the cumulative risk corresponding to the quantile loss function defined by Koenker and Bassett (1978). This led to asymptotically correct confidence intervals.

In this thesis, we propose to adapt this approach to the batch setting and provide nonasymptotic results. We also apply these results to build quarterly prediction and confidence regions for the French Gross Domestic Product (GDP) growth. Our approach is the following. We assume that we are given a set of basic predictors - this is a usual approach in statistical learning, the predictors are sometimes referred as "experts", e.g. Cesa-Bianchi and Lugosi (2006). Following Alquier and Wintenberger (2012), we describe a procedure of aggregation, usually referred as Exponentially Weigthed Agregate (EWA), Dalalyan and Tsybakov (2008); Gerchinovitz (2011), or Gibbs estimator, Catoni (2004, 2007). It is interesting to note that this procedure is also related to aggregations procedure in online learning as the weighted majority algorithm of Littlestone and Warmuth (1994), see also Vovk (1990). We give a PAC-Bayesian inequality that ensures optimality properties for this procedure. In a few words, this inequality claims that our predictor performs as well as the best basic predictor up to a remainder of the order $\mathcal{K}/\sqrt{n}$ where n is the number of observations and $\mathcal{K}$ measures the complexity of the set of basic predictors. This result is very general, two conditions will be required: the time series must be weakly dependent in a sense that we will make more precise below, and loss function must be Lipschitz. This includes, in particular, the quantile loss functions. This allows us to apply this result to our problem of economic forecasting. Under additional assumptions, we are able to prove that the empirical risk minimizer (ERM, see e.g. Vapnik (1999)) is also able to perform such a prediction. Our main results are given

under the form of PAC-Bayesian oracle inequalities.

The idea of PAC-bayesian learning theorems, as introduced by Shawe-Taylor and Williamson (1997); McAllester (1999) is to measure the complexity of models, and thereby their ability to generalize from observed examples to unknown situations, with the help of some prior probability measure defined on the parameter space. "PAC" is fundamentally about choosing particular prediction functions out of some class of plausible alternatives so that, with high reliability, the resulting predictions will be nearly as accurate as possible ("probably approximately correct"). Bayesian analysis of generalization can place a prior distribution on the hypotheses and estimate the volume of the space that is consistent with the training data. The larger this volume the greater the confidence in the classifier obtained. The key feature of such estimators is that they provide a posteriori estimates of the generalization based on properties of the hypothesis and the training data. This contrasts with a 'classical' PAC analysis which provides only a priori bounds. Here, we use for simplicity the term parameter space in a rather loose and unusual way, to talk about the union of all the parameters of all the models we envision. (maybe the term model space would be more accurate : these parameters may be of finite or infinite dimension and we do not restrict the number of models, therefore we are definitely not describing a parametric statistical framework, but rather a non-parametric one!).

The status of the prior measure has not to be misunderstood either : it does not represent the frequency according to which we expect to observe data produced by different probability distributions, nor does it stand for the belief we put in the accuracy of different possible distributions or different possible models. It is somehow equivalent to the choice of some representation of the parameter space, and therefore is related to the Minimum Description Length approach of Rissanen and to the structural risk minimization approach of Vapnik. On a more technical level, it is meant to produce non asymptotic worst case bounds. (as opposed to a Bayesian study of the mean risk under the prior).

In particular, these methods control the expected accuracy of future predictions from mis-specified models based on finite samples. This allows for immediate model comparisons which neither appeal to asymptotic nor make strong assumptions about the data-generating process, in stark contrast to such popular model-selection tools as AIC.

5.2 The context

Let us assume that we observe $X_1, \ldots, X_n$ from a $\mathbb{R}^p$-valued stationary time series $X = (X_t)_{t\in\mathbb{Z}}$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^p$. Fix an integer k and let us assume that we are given a family of predictors

$$\left\{f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta\right\}$$

for any θ and any t, f_θ applied to the last past values $(X_{t-1}, \ldots, X_{t-k})$ is a possible prediction of X_t. The aim is to choose some predictors f_θ which predicts X_t from $(X_{t-1}, \ldots, X_{t-k})$ making as few mistakes as possible on average.

For the sake of simplicity, let us put for any $t \in \mathbb{Z}$ and any $\theta \in \Theta$,

$$\hat{X}_t^\theta = f_\theta(X_{t-1}, \ldots, X_{t-k}).$$

We also assume that $\theta \mapsto f_\theta$ is linear. As we have already explained, the set of predictors $\left\{f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta\right\}$ will in general not be a single parametric model, but rather the union of a large number of parametric models. Using the terminology of statistics, note that we may want to include parametric set of predictors as well as non-parametric ones (i.e. respectively finite dimensional and infinite dimensional).

Example 5.2.1. *We put $\theta = (\theta_0, \theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^{k+1}$ and define the linear autoregressive predictors*

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \theta_0 + \sum_{j=1}^{k} \theta_j X_{t-j}.$$

In order to deal with various family of predictors, we will sometimes use a model-selection type approach:

$$\Theta = \cup_{j=1}^{M} \Theta_j.$$

Example 5.2.2. *We may generalize the previous example to non-parametric auto-regression, for example using a dictionary of functions $(\mathbb{R}^p)^k \to \mathbb{R}^p$, say $(\varphi_i)_{i=0}^{\infty}$. Then we take $\theta = (\theta_1, \ldots, \theta_\ell) \in \Theta_j \subset \mathbb{R}^j$ and*

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \sum_{i=1}^{j} \theta_i \varphi_i(X_{t-1}, \ldots, X_{t-k}).$$

In many cases, $\Theta = \cup_{j=1}^{M} \Theta_j$ will be a finite (or more generally countable) union of subspaces. The importance of introducing such structure has been put forward by V. Vapnik (Vapnik (1999)), as a way to avoid making strong hypotheses on the distribution of the sample.

From the technical point of view, our aim will be to produce non asymptotic bounds for the risk of properly designed predictors of X_t given $(X_{t-1}, \ldots, X_{t-k})$, leading to a non asymptotic level of confidence for this risk.

Come back to the prediction problem, in order to quantifier the prediction $\hat{X}_t^{\theta}$, we first define a quantitative criterion to evaluate the quality of the predictions. Let ℓ be a loss function, the risk of f_θ will be measured as its expected error rate:

Definition 5.2.1. *We put, for any $\theta \in \Theta$,*

$$R(\theta) = \mathbb{E}\left[\ell\left(\hat{X}_t^{\theta}, X_t\right)\right].$$

with $\mathbb{E}$ the expected value of all the observations $(X_t)_{1 \leq t \leq n}$ from a stationary process (X_t).

Note that because of the stationarity, $R(\theta)$ does not depend on t. To actually calculate the risk, we would need to know the distribution of the process $(X_t)_{t \in \mathbb{Z}}$ and have a single fixed prediction function f_θ, neither of which is common. Because explicitly calculating the risk is infeasible, forecasters typically try to estimate it, which calls for detailed assumptions on the distribution. The alternative we employ here is to find upper bounds on risk which hold uniformly over large classes of models Θ from which some particular θ is chosen, possibly in a data dependent way, and uniformly over distributions.

As the above quantity is unobserved, we use the corresponding empirical error rate.

Definition 5.2.2. *For any $\theta \in \Theta$,*

$$r_n(\theta) = \frac{1}{n-k} \sum_{i=k+1}^{n} \ell\left(\hat{X}_i^{\theta}, X_i\right).$$

We cannot minimize $R(\theta)$ with respect to θ because $R(\theta)$ is not observable: it depends on the unknown distribution. The next sensible attempt is to minimize

$r_n(\theta)$ instead. Unfortunately, although $\mathbb{E}(r_n(\theta)) = R(\theta)$, the fluctuations of the random process $r_n(\theta)$ may be strong enough to make the solutions of the two minimization problem quite difficult, and even in many cases completely unrelated. An intensively studied way to get some control on this situation is to add a penalty term $pen(\theta)$ and study the relations between $\inf_\theta R(\theta) + pen(\theta)$ and $\inf_\theta r(\theta) + pen(\theta)$. The penalty $pen(\theta)$ has a regularizing effect: it shrinks the size of the set of values of θ where $\inf_\theta r(\theta) + pen(\theta)$ is likely to be achieved and therefore provides a way to control the gap between $\mathbb{P}[\inf_\theta r(\theta) + pen(\theta)]$ and $\inf_\theta R(\theta) + pen(\theta)$.

5.3 Basic inequality

Let $\mathcal{T}$ be a σ-algebra on Θ and $\mathcal{T}_\ell$ be its restriction to Θ_ℓ. Let $\mathcal{M}_+^1(\Theta)$ denote the set of all probability measures on $(\Theta, \mathcal{T})$, and $\pi \in \mathcal{M}_+^1(\Theta)$. This probability measure is usually called the *prior*. It will be used to control the complexity of the set of predictors Θ.

Note that in statistical learning, given an estimator $\hat{\theta}$, the bounds on the risk $R(\hat{\theta})$ often depends on the empirical risk $r_n(\hat{\theta})$ and on a remainder term measuring the complexity of the model of Θ. The aim of the PAC-Bayesian approach is to obtain PAC bounds on the integrated risk

$$\int_\Theta R(\theta)\rho(\mathrm{d}\theta) = \rho[R(\cdot)]$$

where $\rho \in \mathcal{M}_+^1(\Theta)$ is whatever posterior distribution, depending on π and on the observed data. The bounds here will depend on the empirical counterpart of $\rho[R(\cdot)]$:

$$\rho[r_n(\cdot)] = \int_\Theta r_n(\theta)\rho(\mathrm{d}\theta),$$

and on a measure of the distance between ρ and π. This measure of the distance between ρ and π will be made by the use of the Kullback divergence.

Definition 5.3.1. *The Kullback-Leibler divergence $\mathcal{K}(\rho, \pi)$ of ρ with respect to π is defined as:*

$$\mathcal{K}(\rho, \pi) = \begin{cases} \int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}\rho, & \textit{when } \rho \textit{ is absolutely continuous} \\ & \textit{with respect to } \pi \textit{ i.e } \rho \ll \pi \\ \infty, & \textit{otherwise} \end{cases}$$

The following lemma shows in which sense the Kullback divergence function can be thought of as the dual of the Legendre transform.

Definition 5.3.2. *For any measurable function $h : \Theta \to \mathbb{R}$, for any measure $\rho \in \mathcal{M}_+^1(\Theta)$ we put:*

$$\rho[h(\theta)] = \sup_{B \in \mathbb{R}} \mathbb{E}\left(\min\{B, h(\theta)\}\right)$$

Lemma 5.3.1 (Legendre transform of the Kullback divergence function)**.** *For any $\pi \in \mathcal{M}_+^1(\Theta)$, for any measurable function $h : \Theta \to \mathbb{R}$ we have:*

$$\pi[\exp(h)] = \exp\left(\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left(\rho[h] - \mathcal{K}(\rho, \pi)\right)\right)$$

where (as in general we will note)

$$\pi[h] = \int h(x)\pi(\mathrm{d}x)$$

with convention $\infty - \infty = -\infty$. Indeed, a priority is given to ∞ in ambiguous cases: the expectation of a function whose negative part is not integrable will be assumed to be $-\infty$, even when its positive part integrates to $+\infty$. Moreover, as soon as h is upper-bounded on the support of π, the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure $\pi\{h\}$.

Actually, it seems that in the case of discrete probabilities, this result was already known by Kullback (Problem 8.28 of Chapter 2 in Kullback (1959)). Here we provide a complete proof of this variational formula from Catoni (2003).

Proof. Let us assume that h is upper-bounded on the support of π. Consider the Gibbs distribution, $\pi_{\exp(h)}$ given by:

$$\frac{\mathrm{d}\pi_{\exp(h)}}{\mathrm{d}\pi}(\theta) = \frac{\exp[h(\theta)]}{\pi[\exp[h(\theta)]]},$$

Let us remark that ρ is absolutely continuous with respect to π if and only if it is absolutely continuous with respect to $\pi_{\exp(h)}$. Let us assume that this is the case, then we have,

$$\mathcal{K}(\rho, \pi_{\exp(h)}) = \log\left\{\pi[\exp[h(\theta)]]\right\} + \mathcal{K}(\rho, \pi) - \rho[h(\theta)]$$

The left-hand side of this equation is nonnegative and cancels only for $\rho = \pi_{\exp(h)}$. Note that this equation is still valid if ρ is not absolutely continuous with respect to π. (it just says that $+\infty = +\infty$ in this case). So we obtain:

$$0 = \inf_{\rho \in \mathcal{M}_+^1(\Theta)} [\mathcal{K}(\rho, \pi) - \rho(h)] + \log \pi \exp(h)$$

This proves the second part of lemma 5.3.1, For the first part, we now use the notation $\min\{B, h(\theta)\} = B \wedge h(\theta)$, then we get

$$\begin{aligned}
\log \pi \exp[h(\theta)] &= \sup_{B \in \mathbb{R}} \log \pi \exp[B \wedge h(\theta)] \\
&= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \sup_{B \in \mathbb{R}} \{\rho[B \wedge h(\theta)] - \mathcal{K}(\rho, \pi)\} \\
&= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \sup_{B \in \mathbb{R}} \{\rho[B \wedge h(\theta)]\} - \mathcal{K}(\rho, \pi) \\
&= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[h(\theta)] - \mathcal{K}(\rho, \pi)
\end{aligned}$$

□

We now turn to the study of large deviations for partial sums of weakly dependent processes. Our main tool is Hoeffding type inequalities which provide an upper bound on the probability that the empirical error deviates from its expected value. The aim is to analyze the fluctuations of the random process $\theta \to r_n(\theta)$ from its mean process $\theta \to R(\theta)$. This Hoeffding inequality transform is well suited to relate $\min_{\theta \in \Theta} r_n(\theta)$ to $\inf_{\theta \in \Theta} R(\theta)$, since for large enough values of the parameter λ, corresponding to low enough values of the temperature, the system has small fluctuations around its ground state.

The Hoeffding's inequality is a powerful tool in both probability and statistics. It says that the sum of random variables deviates from its expected value can be upper bounded on the probability. More precisely, when $(X_i)_{1 \leq i \leq n}$ is a sequence of bounded random variables, the Hoeffding-type inequality can be constructed in such a way that

$$\mathbb{E} e^{tf(X_1, \ldots, X_n) - tE(f(X_1, \ldots, X_n))} < e^{nt^2 C}$$

where C is a constant depending on f and X_i.

Example 5.3.1. *Let $X_1, \ldots, X_n$ be i.i.d rondom variables bounded, i.e $a \leq \|X_i\| \leq b$ almost surely. Let $f(x_1, \ldots, x_n) = \sum_{i=1}^n X_i$, we obtain evidently a Hoeffding's inequality with $C = \frac{(b-a)^2}{8}$.*

Let us begin with exponential type inequalities for dependent random variables. Here, we are interested in the $\theta_{\infty,n}(1)$-weak dependence condition of Rio (2000); Dedecker et al. (2007). Let us recall the notation.

Definition 5.3.3. *For any $k > 0$, define the $\theta_{\infty,k}(1)$-weak dependence coefficients of a bounded stationary sequence (X_t) by the relation*

$$\theta_{\infty,k}(1) := \sup_{f \in \Lambda_1^k, 0<j_1<\cdots<j_k} \left\| \mathbb{E}\left[f(X_{j_1}, \ldots, X_{j_\ell})|X_t, t \leq 0\right] - \mathbb{E}\left[f(X_{j_1}, \ldots, X_{j_\ell})\right] \right\|_\infty .$$

where Λ_1^k is the set of 1-Lipshitz functions of q variables

$$\Lambda_1^k = \left\{ f : (\mathbb{R}^p)^k \to \mathbb{R}, \quad \frac{|f(u_1, \ldots, u_k) - f(u_1', \ldots, u_k')|}{\sum_{j=1}^k \|u_j - u_j'\|} \leq 1 \right\}.$$

The sequence $(\theta_{\infty,k}(1))_{k>0}$ is non decreasing and upper bounded for many bounded stationary time series. This notion of dependence is more general for bounded time series than mixing ones, see Dedecker et al. (2007) for details.

Lemma 5.3.2 (Rio (2000))**.** *Let h be a function $(\mathbb{R}^p)^n \to \mathbb{R}$ such that for all x_1, ..., x_n, y_1, ..., $y_n \in \mathbb{R}^p$,*

$$|h(x_1, \ldots, x_n) - h(y_1, \ldots, y_n)| \leq \sum_{i=1}^n \|x_i - y_i\|. \tag{5.1}$$

Let the stationary sequence (X_t) be θ dependent and bounded, i.e. $(\theta_{\infty,k}(1))_{k>0} < \mathcal{C}$ and $\|X_0\| \leq \mathcal{B}$ almost surely. Then for any $t > 0$ we have $C = \frac{(\mathcal{B}+\theta_{\infty,n}(1))^2}{2}$ i.e

$$\mathbb{E}\left(e^{t\{\mathbb{E}[h(X_1,\ldots,X_n)] - h(X_1,\ldots,X_n)\}}\right) \leq e^{\frac{t^2 n(\mathcal{B}+\theta_{\infty,n}(1))^2}{2}}.$$

Proof. This version of Theorem 1 of Rio (2000) comes rewriting the inequality 3 in Rio (2000) as, for any 1-Lipschtiz function g:

$$\Gamma(g) = \|\mathbb{E}\left(g(X_{l+1}, \ldots, X_n)|\mathcal{F}_l\right) - \mathbb{E}\left(g(X_{l+1}, \ldots, X_n)\right)\|_\infty \leq \theta_{\infty,n-l}(1).$$

The result is proved as $\sup_{1 \leq r \leq n} \theta_{\infty,r}(1) \leq \theta_{\infty,n}(1)$. □

Others exponential inequalities can be used to obtain PAC-Bounds in the context of time series: the inequalities in Doukhan (1994); Samson (2000) for mixing time series, and Dedecker et al. (2007); Wintenberger (2010) under weakest "weak dependence" assumptions, Seldin et al. (2012) for martingales. However, Lemma 5.3.2 is particularly convenient here, and will lead to optimal learning rates. In order to prove the fast rates oracle inequalities, a more restrictive dependence condition is assumed. It holds on the uniform mixing coefficients ϕ introduced by Ibragimov (1962).

Definition 5.3.4. *The ϕ-mixing coefficients of the stationary sequence (X_t) with distribution $\mathbb{P}$ are defined as*

$$\phi_r = \sup_{(A,B)\in\sigma(X_t,t\leq 0)\times\sigma(X_t,t\geq r)} |\mathbb{P}(B/A) - \mathbb{P}(B)|.$$

The stationary sequence (X_t) is uniformly mixing when $\phi_r \to 0$. Examples of uniformly mixing sequences are given in Doukhan (1994).

We will also use Samson Berstein's type inequality in the proof of the fast rates.

Lemma 5.3.3 (Samson (2000))**.** *Let $N \in \mathbb{N}$. Let $(Z_i)_{i\in\mathbb{Z}}$ be a stationary process, let (ϕ_r^Z) denote its ϕ-mixing coefficients, let f be a measurable function $\mathbb{R} \to [-M, M]$ and let*

$$S_N(f) := \sum_{i=1}^{N} f(Z_i).$$

Then: we have one inequality with $C = 8K_{\phi^Z}\sigma^2(f)$. i.e.

$$\ln \mathbb{E}(\exp(\lambda(S(f) - \mathbb{E}S(f)))) \leq 8K_{\phi^Z} N\sigma^2(f)\lambda^2$$

for all $0 \leq \lambda \leq 1/(MK_{\phi^Z}^2)$, where $K_{\phi^Z} = 1+\sum_{r=1}^{N}\sqrt{\phi_r^Z}$ and $\sigma^2(f) = \mathrm{Var}\,[f(Z_i)]$.

Proof. Actually, this result is not stated in this form in Samson (2000) but can be deduced from the proof of Theorem 3 in that paper, a much more difficult result. To do so, in page 457 of Samson (2000), just replace the definition of $f_N(x_1, \ldots, x_n)$ by $f_N(x_1, \ldots, x_n) = \sum_{i=1}^{n} g(x_i)$ (following the notations of Samson (2000)). Then check that all the arguments of the proof remain valid, the claim of Lemma 5.3.3 is obtained page 460, line 7. □

5.4 ERM and Gibbs estimator

As the objective is to minimize the risk $R(\cdot)$, naturally, we first consider the empirical risk $r_n(\cdot)$. The boundedness assumption ensures that it is a good estimator of R.

Definition 5.4.1 (ERM estimator Vapnik (1999))**.** *We define the Empirical Risk Minimizer estimator (ERM) by*

$$\hat{\theta}^{ERM} \in \arg\min_{\theta\in\Theta} r_n(\theta).$$

The random measures depending on the empirical risk $r(\theta)$ are a special case of posterior distributions. More precisely, we will make a heavy use of Gibbs estimator distributions of the form:

Definition 5.4.2. *For any measure π and any measurable function h such that $\pi[\exp(h)] < +\infty$, the Gibbs measure denote ρ is defined by*

$$\rho(\mathrm{d}\theta) = \frac{\exp(h(\theta))\pi(\mathrm{d}\theta)}{\int \exp(h(\theta'))\pi(\mathrm{d}\theta')}.$$

The introduction of these posterior distributions, viewed as random objects whose fluctuations are easily manageable, leads us to consider randomized estimators: instead of picking some parameter $\hat{\theta}$ as a deterministic function of the observations $(X_1, \ldots, X_n)$, we choose it at random according to the posterior distribution ρ (which itself depends on the observations).

Remark 5.4.1. *In the case where $\Theta = \cup_{j=1}^{M}\Theta_j$ and the Θ_j are disjoint, we can write*

$$\pi(\mathrm{d}\theta) = \sum_{j=1}^{m} \mu_j \pi_j(\mathrm{d}\theta)$$

where $\mu_j := \pi(\Theta_j)$ and $\pi_j(\mathrm{d}\theta) := \pi(\mathrm{d}\theta)\mathbb{1}_{\Theta_j}(\theta)/\mu_j$. Here π_j can be interpreted as a prior probability measure inside the model Θ_j and that the μ_j as a prior probability measure between the models.

Definition 5.4.3 (Gibbs estimator)**.** *We put, for any $\lambda > 0$,*

$$\hat{\theta}_\lambda = \int_\Theta \theta \hat{\rho}_\lambda(\mathrm{d}\theta)$$

where

$$\hat{\rho}_\lambda(\mathrm{d}\theta) = \frac{e^{-\lambda r_n(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda r_n(\theta')}\pi(\mathrm{d}\theta')}.$$

The choice of the parameter λ is discussed in the next sections. The Gibbs estimator is a method to aggregate estimators who:

- build a posterior distribution which is faster to compute,
- build efficient posterior distributions in the case of a continuous family of fixed distributions, thus avoiding the use of sample splitting schemes.

Our results assert that the risk of the ERM or Gibbs estimator is close to $\inf_\theta R(\theta)$ up to a remainder term Δ called the rate of convergence. For the sake of simplicity, let $\overline{\theta} \in \Theta$ be such that

$$R(\overline{\theta}) = \inf_\theta R(\theta).$$

If $\overline{\theta}$ does not exist, it is replaced by an approximative minimizer $\overline{\theta}_\alpha$ satisfying $R(\overline{\theta}_\alpha) \leq \inf_\theta R(\theta) + \alpha$ where α is negligible w.r.t. Δ (e.g. $\alpha < 1/n$).

We want to prove that the ERM satisfies, for any $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\hat{\theta}^{ERM}\right) \leq R(\overline{\theta}) + \Delta(n, \Theta, \varepsilon)\right) \geq 1 - \varepsilon \tag{5.2}$$

where $\Delta(n, \Theta, \varepsilon) \to 0$ as $n \to \infty$.

We also want to prove that and that the Gibbs estimator satisfies, for any $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\hat{\theta}_\lambda\right) \leq R(\overline{\theta}) + \Delta(n, \lambda, \pi, \varepsilon)\right) \geq 1 - \varepsilon \tag{5.3}$$

where $\Delta(n, \lambda, \pi, \varepsilon) \to 0$ as $n \to \infty$ for some $\lambda = \lambda(n)$. To obtain such results called *oracle inequalities*, we require some general assumptions discussed later.

5.5 Main assumptions and main tools

In order to ensure good performances in terms of prediction for the ERM and Gibbs estimator, we need some hypotheses in the model. Assumptions **LipLoss**(K) and **Lip**(L) hold respectively on the loss function ℓ and the set of predictors Θ. In some extent, we choose the loss function and the predictors, so these assumptions can always be satisfied. Note that assumption **Margin**$(\mathcal{K})$ holds on ℓ and also on the marginal distribution. It is used to obtain fast rates of convergence only and thus we discuss it in Section 5.7. On the other hand, assumptions

WeakDep($\mathcal{C}$) and **PhiMix**($\mathcal{C}$) hold on the dependence of the time series. In practice, we cannot know wether these assumptions are satisfied on data. However, remark that these assumptions are not parametric and are satisfied for many classical models, see Doukhan (1994); Dedecker et al. (2007).

Assumption LipLoss(K), $K > 0$: the loss function ℓ is given by

$$\ell(x, x') = g(x - x')$$

for some convex K-Lipschitz function g such that $g(0) = 0$ and $g \geq 0$.

Example 5.5.1. *A first example is $\ell(x, x') = \|x - x'\|$. In this case, the Lipschitz constant K is 1. This example was studied in detail in Alquier and Wintenberger (2012). In Modha and Masry (1998); Meir (2000), the loss function is the quadratic loss $\ell(x, x') = \|x - x'\|^2$. Note that it also satisfies our Lipschitz condition, but only if we assume that the time series is bounded.*

Example 5.5.2. *When the time-series is real-valued, we can use a quantile loss function. The class of quantile loss functions is defined as*

$$\ell_\tau(x, y) = \begin{cases} \tau (x - y), & \text{if } x - y > 0 \\ -(1 - \tau)(x - y), & \text{otherwise} \end{cases}$$

where $\tau \in (0, 1)$. It is motivated by the following remark: if U is a real-valued random variable, then any value t^* satisfying $\mathbb{P}(U \leq t^*) = \tau$ is a minimizer of of $t \mapsto \mathbb{E}(\ell_\tau(X - t))$; such a value is called quantile of order τ of U. So, the use of this loss function might be a good way to evaluate the risk of rare events and to build confidence intervals. This loss function was introduced by Koenker and Bassett (1978), see Koenker (2005) for a survey. Recently, Belloni and Chernozhukov (2011) used it in the context of high-dimensional regression, and Biau and Patra (2011) in learning problems.

Assumption Lip(L), $L > 0$: for any $\theta \in \Theta$ there are coefficients $a_j(\theta)$ for $1 \leq j \leq k$ such that, for any x_1, ..., x_k and y_1, ..., y_k,

$$\|f_\theta(x_1, \ldots, x_k) - f_\theta(y_1, \ldots, y_k)\| \leq \sum_{j=1}^{k} a_j(\theta) \|x_j - y_j\|,$$

with $\sum_{j=1}^{k} a_j(\theta) \leq L$.

Remark that for bounded observations the empirical risk is a bounded random variable under assumptions **LipLoss**(K) and **Lip**(L). Such condition is required in the approach of individual sequences. We assume it in the statistical approach for simplicity but it is possible to extend the slow rates oracles inequalities to unbounded cases see Alquier and Wintenberger (2012).

Assumption WeakDep$(\mathcal{C})$**,** $\mathcal{C} > 0$: There exists finite constants $(\mathcal{C})$, $\mathcal{C} > 0$, such that $\sup_{t\in\mathbb{Z}} \|X_t\| \leq \mathcal{B}$ almost surely, and $\theta_{\infty,k}(1) \leq \mathcal{C}$ for any $k > 0$.

Under this assumption, the process (X_t) will be called θ weakly dependent.

Example 5.5.3. *Examples of processes satisfying* **WeakDep**$(\mathcal{C})$ *are provided in Alquier and Wintenberger (2012); Dedecker et al. (2007). It includes Bernoulli shifts* $X_t = H(\xi_t, \xi_{t-1}, \xi_{t-2}, \dots)$ *where the* ξ_t *are iid,* $\|\xi_0\| \leq b$ *and* H *satisfies a Lipschitz condition:*

$$\|H(v_1, v_2, ...) - H(v_1', v_2', ...)\| \leq \sum_{j=0}^{\infty} a_j \|v_j - v_j'\| \text{ with } \sum_{j=0}^{\infty} j a_j < \infty.$$

Then (X_t) *is bounded by* $\mathcal{B} = H(0, 0, ...) + b\mathcal{C}$ *and satisfies* **WeakDep**$(\mathcal{C})$ *with* $\mathcal{C} = \sum_{j=0}^{\infty} j a_j$*. In particular, solutions of linear* ARMA *models with bounded innovations satisfy* **WeakDep**$(\mathcal{C})$*.*

In order to prove the fast rates oracle inequalities, a more restrictive dependence condition is assumed.

Assumption PhiMix$(\mathcal{C}')$**,** $\mathcal{C}' > 0$: $1 + \sum_{r=1}^{\infty} \sqrt{\phi_r} \leq \mathcal{C}'$.

This assumption is more restrictive than **WeakDep**$(\mathcal{C})$ for bounded time series:

Proposition 5.5.1 (Rio (2000))**.** *For bounded time series,* **PhiMix**$(\mathcal{C}') \Rightarrow$ **WeakDep**$(\mathcal{C})$.

For the sake of completeness, we give the proof of this already known result.

Proof. First let us remind

$$\theta_{\infty,n}(1) \leq \sum_{i=1}^{n} \|\mathbb{E}(\|X_i - X_i^*\| / \sigma(X_t, t \leq 0))\|_{\infty}.$$

Now we will consider the maximal coupling scheme of Goldstein (1979): there exists a version (X_t^*) such that

$$\begin{aligned}&\|\mathbb{P}(X_t \neq X_t^* \text{ for some } \ t \geq r/\sigma(X_t, t \leq 0))\|_\infty \\ &= \sup_{(A,B)\in\, \sigma(X_t,t\leq 0)\times\sigma(X_t,t\geq r)} |\mathbb{P}(B/A) - \mathbb{P}(B)| = \phi(r)\end{aligned}$$

We know that, for any variables Y, Z bounded by $\|X_0\|_\infty$, $\|Y - Z\| \leq 2\|X_0\|_\infty \mathbb{1}_{Y\neq Z}$. So

$$\begin{aligned}\|\mathbb{E}(\|X_i - X_i^*\|/\sigma(X_t, t \leq 0))\|_\infty &\leq 2\|X_0\|_\infty \|\mathbb{E}(\mathbb{1}_{X_i\neq X_i^*}/\sigma(X_t, t \leq 0))\|_\infty \\ &\leq 2\|X_0\|_\infty \|\mathbb{P}(X_i \neq X_{i^*}/\sigma(X_t, t \leq 0))\|_\infty \\ &\leq 2\|X_0\|_\infty \|\mathbb{P}(X_t \neq X_t^* \text{ for some } \ t \geq r/\sigma(X_t, t \leq 0))\|_\infty \\ &\leq 2\|X_0\|_\infty \|\phi(r).\end{aligned}$$

We conclude $\theta_{\infty,n}(1) \leq 2\|X_0\|_\infty \| \sum_{r=1}^n \phi(r)$. □

For fast rates oracle inequalities, we use an additional assumption that mix optimal properties of the loss function ℓ and the margin distributions. In the iid case, such conditions are also required. They are called Margin assumptions Mammen and Tsybakov (1999); Alquier (2008) or Bernstein hypothesis Lecué (2011).

Assumption Margin($\mathcal{K}$), $\mathcal{K} > 0$:

$$\mathbb{E}_{\mathbb{P}}\left\{\left[\ell\Big(X_{q+1}, f_\theta(X_q, ..., X_1)\Big) - \ell\Big(X_{q+1}, f_{\overline{\theta}}(X_q, ..., X_1)\Big)\right]^2\right\} \leq \mathcal{K}\left[R(\theta) - R(\overline{\theta})\right].$$

Theorem 5.5.1 (PAC-Bayesian Oracle Inequality for the Gibbs estimator)**.** *Let us assume that* **LowRates(κ)** *is satisfied for some $\kappa > 0$. Then, for any λ, $\varepsilon > 0$ we have*

$$\mathbb{P}\left\{R\left(\hat{\theta}_\lambda\right) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)}\left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(2/\varepsilon\right)}{\lambda}\right]\right\} \geq 1 - \varepsilon.$$

This result is the analogous of the PAC-Bayesian bounds proved by Catoni in the case of iid data Catoni (2007). It is proved in Section 5.10. This very general result provides a bound on the generalization risk of the Gibbs estimator $\hat{\theta}_\lambda$. Two question arise now:

(1) when one uses a given class of predictor Θ, what is the value of this bound?

(2) what value of λ should be taken in order to minimize this bound?

The next section will provide answers to these questions. Note that we will see that in some particular cases, the ERM $\hat{\theta}^{ERM}$ will predict as well as the Gibbs estimator with optimal parameter λ. So in these cases, the question of the choice of λ vanishes. However, such a general result as Theorem 5.5.1 cannot be proved for the ERM (see Vapnik (1999): we need assumptions on θ).

5.6 Low rates oracle inequalities

In this section, we give oracle inequalities (5.2) and/or (5.3) with low rates of convergence $\Delta(n, \Theta) \sim \sqrt{c(\Theta)/n}$ and also the proof of these results.

5.6.1 Finite classes of predictors

Consider first the toy example where Θ is finite with $|\Theta| = M$, $M \geq 1$. In this case, the optimal rate in the iid case is known to be $\sqrt{\log(M)/n}$, see e.g. Vapnik (1999).

Theorem 5.6.1. *Assume that* $|\Theta| = M$ *and that* **LowRates(κ)** *is satisfied for* $\kappa > 0$. *Let* π *be the uniform probability distribution on* Θ. *Then the oracle inequality* (5.3) *is satisfied for any* $\lambda > 0$, $\varepsilon > 0$ *with*

$$\Delta(n, \lambda, \pi, \varepsilon) = \frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda}.$$

Proof. We apply Theorem 5.5.1 for $\pi = \frac{1}{M}\sum_{\theta\in\Theta}\delta_\theta$ and restrict the inf in the upper bound to Dirac masses $\rho \in \{\delta_\theta, \theta \in \Theta\}$. We obtain $\mathcal{K}(\rho, \pi) = \log M$, and the upper bound for $R(\hat{\theta}_\lambda)$ becomes:

$$R\left(\hat{\theta}_\lambda\right) \leq \inf_{\rho\in\{\delta_\theta, \theta\in\Theta\}} \left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda}\right]$$

$$= \inf_{\theta \in \Theta} \left[R(\theta) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda} \right].$$

□

The choice of λ in practice in this toy example is already not trivial. The choice $\lambda = \sqrt{\log(M)n}$ yields the oracle inequality:

$$R(\hat{\theta}_\lambda) \leq R(\overline{\theta}) + 2\sqrt{\frac{\log(M)}{n}} \left(\frac{\kappa}{1-k/n} \right)^2 + \frac{2\log\left(2/\varepsilon\right)}{\sqrt{n\log(M)}}.$$

However, this choice is not optimal and one would like to choose λ as the minimizer of the upper bound

$$\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(M\right).}{\lambda}$$

However $\kappa = \kappa(K, L, \mathcal{B}, \mathcal{C})$ and the constants $\mathcal{B}$ and $\mathcal{C}$ are, usually, unknown.In this context we will prefer the ERM predictor that performs as well as the Gibbs estimator with optimal λ:

Theorem 5.6.2. *Assume that* $|\Theta| = M$ *and that* **LowRates(κ)** *is satisfied for* $\kappa > 0$.. *Then the oracle inequality* (5.2) *is satisfied for any* $\varepsilon > 0$ *with*

$$\Delta(n, \Theta, \varepsilon) = \inf_{\lambda > 0} \left[\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda} \right] = \frac{4\kappa}{1-k/n}\sqrt{\frac{\log\left(2M/\varepsilon\right)}{n}}.$$

The proof of this result is given in Section 5.10.

5.6.2 Linear autoregressive predictors

We focus on the linear predictors given in Example 5.2.1.

Theorem 5.6.3. *Consider the linear autoregressive model of* AR(k) *predictors*

$$f_\theta(x_{t-1}, \ldots, x_{t-k}) = \theta_0 + \sum_{j=1}^{k} \theta_j x_{t-j}$$

with

$$\theta \in \Theta = \{\theta \in \mathbb{R}^{k+1}, \|\theta\| \leq L\}$$

such that $\mathbf{Lip}(L)$ *is satisfied. Assume that Assumptions* $\mathbf{LipLoss}(K)$ *and* $\mathbf{WeakDep}(\mathcal{C})$ *are satisfied. Let* π *be the uniform probability distribution on the extended parameter set* $\{\theta \in \mathbb{R}^{k+1}, \|\theta\| \leq L+1\}$. *Then the oracle inequality* (5.3) *is satisfied for any* $\lambda > 0$, $\varepsilon > 0$ *with*

$$\Delta(n, \lambda, \pi, \varepsilon) = \frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + 2\frac{(k+1)\log\left(\frac{(K\mathcal{B}\vee K^2\mathcal{B}^2)(L+1)\sqrt{e}\lambda}{k+1}\right) + \log\left(2/\varepsilon\right)}{\lambda}.$$

In theory, λ can be chosen of the order $\sqrt{(k+1)n}$ to achieve the optimal rates $\sqrt{(k+1)/n}$ up to a logarithmic factor. But the choice of the optimal λ in practice is still a problem. The ERM predictor still performs as well as the Gibbs predictor with optimal λ but under an additional necessary constraint on λ:

Theorem 5.6.4. *Under the assumptions of Theorem 5.6.3, the oracle inequality* (5.2) *is satisfied for any* $\varepsilon > 0$ *with*

$$\Delta(n, \Theta, \varepsilon) = \inf_{\lambda \geq 2K\mathcal{B}/(k+1)} \left[\frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + \frac{(k+1)\log\left(\frac{2eK\mathcal{B}(L+1)\lambda}{k+1}\right) + 2\log\left(2/\varepsilon\right)}{\lambda}\right].$$

The additional constraint on λ does not depend on n. It is restrictive only when $k+1$, the complexity of the autoregressive model, has the same order than n. For n sufficiently large and $\lambda = ((1 - k/n)/\kappa)\sqrt{((k+1)n/2)}$ satisfying the constraint $\lambda \geq 2K\mathcal{B}/(k+1)$ we obtain the oracle inequality

$$R(\hat{\theta}^{ERM}) \leq R(\overline{\theta}) + \sqrt{\frac{2(k+1)}{n}}\frac{\kappa}{1 - k/n}\log\left(\frac{2e^2K\mathcal{B}(R+1)}{\kappa}\sqrt{\frac{n}{k+1}}\right) + \frac{2\sqrt{2}\kappa\log\left(2/\varepsilon\right)}{\sqrt{(k+1)n}\left(1 - k/n\right)}.$$

The optimal slow rate of convergence is achieved up to a logarithmic factor. Theorems 5.6.3 and 5.6.4 are both direct consequences of the following results about general classes of predictors.

5.6.3 General parametric classes of predictors

We state a general result about finite-dimensional families of predictors. The complexity $k+1$ of the autoregressive model is replaced by a more general measure of the dimension $d(\Theta, \pi)$. We also introduce some general measure $D(\Theta, \pi)$ of the diameter of the compact model.

Theorem 5.6.5. *Assume that* **LowRates(κ)** *is satisfied and the existence of* $d = d(\Theta, \pi) > 0$ *and* $D = D(\Theta, \pi) > 0$ *satisfying the relation*

$$\forall \delta > 0, \quad \log \frac{1}{\int_{\theta \in \Theta} \mathbb{1}\{R(\theta) - R(\overline{\theta}) < \delta\} \pi(\mathrm{d}\theta)} \leq d \log\left(\frac{D}{\delta}\right).$$

Then the oracle inequality (5.3) *is satisfied for any* $\lambda > 0$, $\varepsilon > 0$ *with*

$$\Delta(n, \lambda, \pi, \varepsilon) = \frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + 2\frac{d \log\left(D\sqrt{e}\lambda/d\right) + \log\left(2/\varepsilon\right)}{\lambda}.$$

We remind that the proofs are given in Section 5.10. A similar result holds for the ERM predictor under a more restrictive assumption on the structure of Θ, see Remark 5.6.1.

Theorem 5.6.6. *Assume that*

1. $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D\}$,
2. $\|\hat{X}_1^{\theta_1} - \hat{X}_2^{\theta_2}\| \leq \psi . \|\theta_1 - \theta_2\|_1$ *a.s. for some* $\psi > 0$ *and all* $(\theta_1, \theta_2) \in \Theta^2$.

Assume also that **LipLoss**(K) *and* **WeakDep**($\mathcal{C}$). *are satisfied and that* **Lip**(L) *holds on the extended model* $\Theta' = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D + 1\}$. *Then the oracle inequality* (5.2) *is satisfied for any* $\varepsilon > 0$ *with*

$$\Delta(n, \Theta, \varepsilon) = \inf_{\lambda \geq 2K\psi/d} \left[\frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + \frac{d \log\left(2eK\psi(D+1)\lambda/d\right) + 2\log\left(2/\varepsilon\right)}{\lambda}\right].$$

The proof of this result can be found in Section 5.10. This result yields to nearly optimal rates of convergence for the ERM predictors. Indeed, for n sufficiently large and $\lambda = ((1 - k/n)/\kappa)\sqrt{(dn/2)} \geq 2K\psi/d$ we obtain the oracle inequality

$$R(\hat{\theta}^{ERM}) \leq R(\overline{\theta}) + \sqrt{\frac{2d}{n}} \frac{\kappa}{1 - k/n} \log\left(\frac{2e^2 K\psi(D+1)}{\kappa}\sqrt{\frac{n}{d}}\right) + \frac{2\sqrt{2}\kappa \log\left(2/\varepsilon\right)}{\sqrt{dn}\left(1 - k/n\right)}.$$

Thus, the ERM procedure yields prediction that are close to the oracle with an optimal rate of convergence up to a logarithmic factor. Note that the context of Theorem 5.6.6 are less general than the one of Theorem 5.6.5:

Remark 5.6.1. *Under the assumptions of Theorem 5.6.6 we have for any* $\theta \in \Theta$

$$R(\theta) - R(\overline{\theta}) = \mathbb{E}\left\{g\left(\hat{X}_1^{\theta} - X_1\right) - g\left(\hat{X}_1^{\overline{\theta}} - X_1\right)\right\}$$

$$\leq \mathbb{E}\left\{K\left\|\hat{X}_1^{\theta} - \hat{X}_1^{\overline{\theta}}\right\|\right\}$$
$$\leq K\psi\|\theta - \overline{\theta}\|_1.$$

Define π *as the uniform distribution on* $\Theta' = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D+1\}$. *We derive from simple computation the inequality*

$$\log \frac{1}{\int_{\theta\in\Theta} \mathbb{1}\{R(\theta) - R(\overline{\theta}) < \delta\}\pi(\mathrm{d}\theta)} \leq \log \frac{1}{\int_{\theta\in\Theta} \mathbb{1}\{\|\theta - \overline{\theta}\|_1 < \frac{\delta}{K\psi}\}\pi(\mathrm{d}\theta)}$$
$$\begin{cases} = d\log\left(\frac{K\psi(D+1)}{\delta}\right) \ \textit{when } \delta/K\psi \leq 1 \\ \\ \leq d\log\left(K\psi(D+1)\right) \ \textit{otherwise.} \end{cases}$$

Thus, in any case,

$$\log \frac{1}{\int_{\theta\in\Theta} \mathbb{1}\{R(\theta) - R(\overline{\theta}) < \delta\}\pi(\mathrm{d}\theta)} \leq d\log\left(\frac{(K\psi \vee K^2\psi^2)(D+1)}{\delta}\right)$$

and the assumptions of Theorem 5.6.5 are satisfied for $d(\Theta,\pi) = d$ *and* $D(\Theta,\pi) = (K\psi \vee K^2\psi^2)(D+1)$.

5.6.4 Aggregation in the model-selection setting

Consider now several models of predictors Θ_1, ..., Θ_M and consider

$$\Theta = \bigsqcup_{i=1}^{M} \Theta_i$$

(disjoint union). Our aim is to predict as well as the best predictors among all Θ_j's, but paying only the price for learning in the smallest possible Θ_j. For this, let us choose M priors π_j on each models such that $\pi_j(\Theta_j) = 1$ for all $j \in \{1,...,M\}$. Let $\pi = \sum_{j=1}^{M} p_j\pi_j$ be a mixture of these priors with prior weights $p_j \geq 0$ satisfying $\sum_{j=1}^{M} p_j = 1$. Denote

$$\overline{\theta}_j \in \arg\min_{\theta\in\Theta_j} R(\theta)$$

the oracle of the model Θ_j for any $1 \leq j \leq M$. For any $\lambda > 0$, denote $\hat{\rho}_{\lambda,j}$ the Gibbs distribution on Θ_j and

$$\hat{\theta}_{\lambda,j} = \int_{\Theta_j} \theta\hat{\rho}_{\lambda,j}(\mathrm{d}\theta)$$

the corresponding Gibbs estimator. A Gibbs predictor based on a model selection procedure satisfies an oracle inequality with low rate of convergence:

Theorem 5.6.7. *Assume that:*

1. **LipLoss**(K) *is satisfied for some* $K > 0$*;*

2. **WeakDep**$(\mathcal{C})$ *is satisfied for some* $\mathcal{C} > 0$*;*

3. *for any* $j \in \{1, ..., M\}$ *we have*

   (a) **Lip**(L_j) *is satisfied by the model* Θ_j *for some* $L_j > 0$*,*

   (b) *there are constants* $d_j = d(\Theta_j, \pi)$ *and* $D_j = c(\Theta_j, \pi_j)$ *are such that*

$$\forall \delta > 0, \quad \log \frac{1}{\int_{\theta \in \Theta_j} \mathbb{1}\{R(\theta) - R(\overline{\theta}_j) < \delta\} \pi_j(\mathrm{d}\theta)} \leq d_j \log\left(\frac{D_j}{\delta}\right)$$

Denote $\kappa_j = \kappa(K, L_j, \mathcal{B}, \mathcal{C}) = K(1 + L_j)(\mathcal{B} + \mathcal{C})/\sqrt{2}$ *and define* $\hat{\theta} = \hat{\theta}_{\lambda_{\hat{j}}, \hat{j}}$ *where*

$$\hat{j} = \arg\min_{1 \leq j \leq M} \left\{ \int_{\Theta_j} r_n(\theta) \hat{\rho}_{\lambda_j, j}(\mathrm{d}\theta) + \frac{\lambda_j \kappa_j}{n(1 - k/n)^2} + \frac{\mathcal{K}(\hat{\rho}_{\lambda_j, j}, \pi_j) + \log\left(2/(\varepsilon p_j)\right)}{\lambda_j} \right\}$$

with

$$\lambda_j = \arg\min_{\lambda > 0} \left[\frac{2\lambda \kappa_j^2}{n\left(1 - k/n\right)^2} + 2\frac{d_j \log\left(D_j e \lambda / d_j\right) + \log\left(2/(\varepsilon p_j)\right)}{\lambda} \right].$$

Then, with probability at least $1 - \varepsilon$*, the following oracle inequality holds*

$$R(\hat{\theta}) \leq \inf_{1 \leq j \leq M} \left[R(\overline{\theta}_j) + 2\frac{\kappa_j}{1 - k/n} \left\{ \sqrt{\frac{d_j}{n}} \log\left(\frac{D_j e^2}{\kappa_j} \sqrt{\frac{n}{d_j}} \right) + \frac{\log\left(2/(\varepsilon p_j)\right)}{\sqrt{n d_j}} \right\} \right].$$

The proof of this result is given in 5.10. A similar result can be obtained if we replace the Gibbs predictor in each model by the ERM predictor in each model. The resulting procedure is known in the iid case under the name SRM (Structural Risk Minimization), see Vapnik (1999), or penalized risk minimization, **?**. However, as it was already the case for a fixed model, additional assumptions are required to deal with ERM predictors. In the model-selection context, the procedure to choose among all the ERM predictors also depends on the unknown κ_j's. Thus the model-selection procedure based on Gibbs predictors outperforms the one based on the ERM predictors.

5.7 Fast rates oracle inequalities

5.7.1 Discussion on the assumptions

In this section, we study conditions under which the rate $1/n$ can be achieved. These conditions are restrictive:

- now $p = 1$, i.e. the process $(X_t)_{t\in\mathbb{Z}}$ is real-valued;
- the dependence condition $\mathbf{WeakDep}(\mathcal{C})$ is replaced by $\mathbf{PhiMix}(\mathcal{C})$;
- we assume additionally $\mathbf{Margin}(\mathcal{K})$ for some $\mathcal{K} > 0$.

Let us provide some examples of processes satisfying the uniform mixing assumption $\mathbf{PhiMix}(\mathcal{B}, \mathcal{C})$. In the three following examples (ϵ_t) denotes an iid sequence (called the innovations).

Example 5.7.1 (AR(p) process)**.** *Consider the stationary solution (X_t) of an AR(p) model: $\forall t \in \mathbb{Z}$, $X_t = \sum_{j=1}^{p} a_j X_{t-j} + \epsilon_t$. Assume that (ϵ_t) is bounded with a distribution possessing an absolutely continuous component. If $\mathcal{A}(z) = \sum_{j=1}^{p} a_j z^j$ has no root inside the unit disk in $\mathbb{C}$ then (X_t) is a geometrically ϕ-mixing processe, see Athreya and Pantula (1986) and $\mathbf{PhiMix}(\mathcal{C})$ is satisfied for some $\mathcal{C}$.*

Example 5.7.2 (MA(p) process)**.** *Consider the stationary process (X_t) such that $X_t = \sum_{j=1}^{p} b_j \epsilon_{t-j}$ for all $t \in \mathbb{Z}$. By definition, the process (X_t) is stationary and ϕ-dependent - it is even p-dependent, in the sense that $\phi_r = 0$ for $r > p$. Thus $\mathbf{PhiMix}(\mathcal{C})$ is satisfied for some $\mathcal{C} > 0$.*

Example 5.7.3 (Non linear processes)**.** *For extensions of the AR(p) model of the form $X_t = F(X_{t-1}, \ldots, X_{t-p}; \epsilon_t)$, Φ-mixing coefficients can also be computed and satisfy $\mathbf{PhiMix}(\mathcal{C})$. See e.g. Meyn and Tweedie (1993).*

We now provide an example of predictive model satisfying all the assumptions required to obtain fast rates oracle inequalities, in particular $\mathbf{Margin}(\mathcal{K})$, when the loss function ℓ is quadratic, i.e. $\ell(x, x') = (x - x')^2$:

Example 5.7.4. *Consider Example 5.2.2 where*

$$f_\theta(X_{t-1},\dots,X_{t-k}) = \sum_{i=1}^{N} \theta_i \varphi_i(X_{t-1},\dots,X_{t-k}),$$

for functions $(\varphi_i)_{i=0}^{\infty}$ *of* $(\mathbb{R}^p)^k$ *to* $\mathbb{R}^p$*, and* $\theta = (\theta_1,\dots,\theta_{\mathbb{N}}) \in \mathbb{R}^N$*. Assume the* φ_i *upper bounded by a constant* Φ *and* $\Theta = \{\theta \in \mathbb{R}^N, \|\theta\|_1 \leq D\}$ *such that* $\mathbf{Lip}(L)$ *is satisfied for* $L = D\Phi$*. Moreover* $\mathbf{LipLoss}(K)$ *is satisfied with* $K = 4\mathcal{B}$*. Assume that* $\overline{\theta} = \arg\min_{\theta\in\mathbb{R}^N} R(\theta) \in \Theta$ *in order to have:*

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\left\{\left[\left(X_{q+1} - f_\theta(X_q,...,X_1)\right)^2 - \left(X_{q+1} - f_{\overline{\theta}}(X_q,...,X_1)\right)^2\right]^2\right\} \\
&= \mathbb{E}_{\mathbb{P}}\Big\{\left[f_\theta(X_q,...,X_1) - f_{\overline{\theta}}(X_q,...,X_1)\right]^2 \\
&\qquad\qquad \left[2X_{q+1} - f_\theta(X_q,...,X_1) - f_{\overline{\theta}}(X_q,...,X_1)\right]^2\Big\} \\
&\leq \mathbb{E}_{\mathbb{P}}\left\{\left[f_\theta(X_q,...,X_1) - f_{\overline{\theta}}(X_q,...,X_1)\right]^2 4\mathcal{B}^2(1+R)^2\right\} \\
&\leq 4\mathcal{B}^2(1+R)^2\left[R(\theta) - R(\overline{\theta})\right] \text{ by Pythagorean theorem.}
\end{aligned}$$

Assumption $\mathbf{Margin}(\mathcal{K})$ *is satisfied with* $\mathcal{K} = 4\mathcal{B}^2(1+D)^2$ *and the oracle inequality with fast rates holds if Assumption* $\mathbf{PhiMix}(\mathcal{C})$ *is satisfied.*

5.7.2 General result

We only give oracle inequalities for the Gibbs predictor in the model-selection setting. In the case of one single model, this result can be extended to the ERM predictor. For several models, the approach based on the ERM predictors requires a penalized risk minimization procedure as in the slow rates case. In the fast rates case, the Gibbs predictor itself directly have nice properties. Let $\Theta = \bigsqcup_{i=1}^{M} \Theta_i$ (disjoint union), choose $\pi = \sum_{j=1}^{M} p_j \pi_j$ and denote $\overline{\theta}_j \in \arg\min_{\theta\in\Theta_j} R(\theta)$ as previously.

Theorem 5.7.1. *Assume that:*

1. $\mathbf{Margin}(\mathcal{K})$ *and* $\mathbf{LipLoss}(K)$ *are satisfied for some* K*,* $\mathcal{K} > 0$*;*
2. $\mathbf{PhiMix}(\mathcal{B},\mathcal{C})$ *is satisfied for some* $\mathcal{C} > 0$*;*
3. $\mathbf{Lip}(L)$ *is satisfied for some* $L > 0$*;*

4. for any $j \in \{1, ..., M\}$, *there exist* $d_j = d(\Theta_j, \pi)$ *and* $R_j = R(\Theta_j, \pi_j)$ *satisfying the relation*

$$\forall \delta > 0, \quad \log \frac{1}{\int_{\theta \in \Theta_j} \mathbb{1}\{R(\theta) - R(\overline{\theta}_j) < \delta\} \pi_j(\mathrm{d}\theta)} \leq d_j \log\left(\frac{D_j}{\delta}\right).$$

Then for

$$\lambda = \frac{n-k}{4kKL\mathcal{BC}} \wedge \frac{n-k}{16k\mathcal{C}}$$

the oracle inequality (5.3) *for any* $\varepsilon > 0$ *with*

$$\Delta(n, \lambda, \pi, \varepsilon) = 4 \inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + 4k\mathcal{C}\,(4 \vee KL\mathcal{B}) \frac{d_j \log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{n-k} \right\}.$$

We remind that the proofs are given in Section 5.10. Compare with the low rates case, we don't optimize with respect to λ as the optimal order for λ is independent of j. In practice, the value of λ provided by Theorem 5.7.1 is too conservative. In the iid case, it is shown in Dalalyan and Tsybakov (2008) that the value $\lambda = n/(4\sigma^2)$, where σ^2 is the variance of the noise of the regression yields good results. In our simulations results, we will use $\lambda = n/\hat{\mathrm{var}}(X)$, where $\hat{\mathrm{var}}(X)$ is the empirical variance of the observed time series.

Notice that for the index j_0 such that $R(\overline{\theta}_{j_0}) = R(\overline{\theta})$ we obtain:

$$R\left(\hat{\theta}_\lambda\right) \leq R(\overline{\theta}) + 4k\mathcal{C}\,(4 \vee KL\mathcal{B}) \frac{d_{j_0} \log\left(c_{j_0} e(n-k)/(16k\mathcal{C}d_{j_0})\right) + \log\left(2/(\varepsilon p_{j_0})\right)}{n-k}.$$

So, the oracle inequality achieves the fast rate $d_{j_0}/n \log(n/d_{j_0})$ where j_0 is the model of the oracle. However, note that the choice $j = j_0$ does not necessarily reach the infimum in Theorem 5.7.1.

Let us compare the rates in Theorem 5.7.1 to the ones in Meir (2000); Modha and Masry (1998); Agarwal and Duchi (2011); Agarwal et al. (2012). In Meir (2000); Modha and Masry (1998), the optimal rate $1/n$ is never obtained. The paper Agarwal and Duchi (2011) proves fast rates for online algorithms that are also computationally efficient, see also Agarwal et al. (2012). The fast rate $1/n$ is reached when the coefficients (ϕ_r) are geometrically decreasing. In other cases, the rate is slower. Note that we do not suffer such a restriction. The Gibbs estimator of Theorem 5.7.1 can also be computed efficiently thanks to MCMC procedures, see Alquier and Lounici (2011); Dalalyan and Tsybakov (2008).

5.7.3 Corollary: sparse autoregression

We consider the sparse autoregression model where the number of parameter p is lager than the sample size n. Let the predictors are the linear $\mathrm{AR}(p)$

$$\hat{X}_p^\theta = \sum_{j=1}^{p} X_{p-j}\theta_j.$$

For any $J \subset \{1, \dots, p\}$, define the model:

$$\Theta_J = \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq L \text{ and } \theta_j \neq 0 \Leftrightarrow j \in J\}.$$

Let us remark that we have the disjoint union

$$\Theta = \bigsqcup_{J \subset \{1,\dots,p\}} \Theta_J = \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq 1\}.$$

We choose π_J as the uniform probability measure on Θ_J and $p_j = 2^{-|J|-1}\binom{p}{|J|}^{-1}$.

For any subset $J \subset \{1, \dots, p\}$ define

$$\overline{\theta}_J = \arg\min_{\theta \in \mathbb{R}^p} R(\theta) \in \Theta_J$$

and

$$\overline{\theta} = \arg\min_{\theta \in \mathbb{R}^p} R(\theta) \in \Theta.$$

We can now state the main result for the sparse autoregression.

Corollary 5.7.1. *Assume that* $\mathbf{PhiMix}(\mathcal{C})$ *is satisfied for some* $\mathcal{C} > 0$. *Then the oracle inequality* (5.3) *is satisfied for any* $\varepsilon > 0$ *with*

$$\Delta(n, \lambda, \pi, \varepsilon) = 4 \inf_J \left\{ R(\overline{\theta}_J) - R(\overline{\theta}) + \mathrm{cst.} \frac{|J| \log\left((n-k)p/|J|\right) + \log\left(\frac{2}{\varepsilon}\right)}{n-k} \right\}$$

for some constant $\mathrm{cst} = \mathrm{cst}(\mathcal{B}, \mathcal{C}, L)$.

This extends the results of Alquier and Lounici (2011); Dalalyan and Tsybakov (2008); Gerchinovitz (2011) to the case of autoregression.

Proof. The proof follows the computations of Example 5.7.4 that we do not reproduce here: we check the conditions $\mathbf{LipLoss}(K)$ with $K = 4\mathcal{B}$, $\mathbf{Lip}(L)$ and $\mathbf{Margin}(\mathcal{K})$ with $\mathcal{K} = 4\mathcal{B}^2(1+L)^2$. We can apply Theorem 5.7.1 with $d_J = |J|$ and $D_j = L$. □

5.8 Application to French GDP forecasting

In this section we give an application of the previous result to French GDP forecasting.

5.8.1 Setting of the Problem: Uncertainty in GDP Forecasting

Every quarter, economic forecasters at INSEE [1] publish a forecast of the quarterly growth rate of the French GDP (Gross Domestic Product). Since it involves a huge amount of data that takes months to be collected and processed, the "true" realization of the GDP growth rate $\log(\mathrm{GDP}_t/\mathrm{GDP}_{t-1})$ is only known after a long time (two years). This means that at time $t+1$, the value $\log(\mathrm{GDP}_t/\mathrm{GDP}_{t-1})$ is actually not known. However, a preliminary value of the growth rate is published 45 days only after the end of the current quarter t. This value is called a *flash estimate* and is the quantity that INSEE forecasters actually try to predict. As we want to play exactly the same "game" as the INSEE, we will now focus on the prediction on the flash estimate and let $\Delta\mathrm{GDP}_t$ denote this quantity. In order to do so, they use two sources of information:

1. past flash estimates [2] $\Delta\mathrm{GDP}_t$;
2. a *climate indicator* I_t based on *business surveys*.

A business survey is a questionnaire of about ten questions sent monthly to a representative panel of French companies (see Devilliers (2004) for more details on this process). As a consequence these surveys provide information coming directly from the true economic decision makers. Morever, they are rapidly available (on a monthly basis). Note that a similar approach is used in other countries, see e.g. Biau et al. (2008) on forecasting the European Union GDP growth thanks to EUROSTATS data.

INSEE publishes a composite indicator, the *French business climate indicator*. This indicator summarises information of the whole business survey. Its defini-

1. *Institut National de la Statistique et des Etudes Economiques*, the French national bureau of statistics, http://www.insee.fr/

2. It has been checked that to replace past flash estimates by the actual GDP growth rate when it becomes available do not improve the quality of the forecasting Minodier (2010).

tion is given for example in Clavel and Minodier (2009); Dubois and Michaux (2006). Let I_t denote this indicator at time t (following Cornec (2010), I_t is the mean of the climate indicator at month 3 of quarter $t-1$ and at month 1 and 2 of quarter t, that *are all available* to INSEE forecasters at quarter t when they publish their forecast of $t+1$) All these values (GDP, climate indicator) are available from the INSEE website.

However it is well known that interval confidence or any relevant information about the accuracy of the prediction should be given with the forecast, in order to provide a quantification of its uncertainty. As a consequence the ASA and the NBER started using density forecasts in 1968, while the Central Bank of England and INSEE provide their prediction with a "fan chart". See Diebold et al. (1997); Tay and Wallis (2007) for surveys on density forecasting in official statistics and Britton et al. (1998) for fan charts. However the methodology used is often very crude, see the criticism in Cornec (2010); Dowd (2004). For example, until 2012, the fan chart provided by the INSEE was based on the assumption that the forecast errors are Gaussian with a constant variance. This led to confidence intervals with constant length. But on the other hand there is an empirical evidence that

1. it is more difficult to forecast GDP in a period of crisis or recession;
2. the distribution of the errors is non-symmetric.

See e.g. the graphics in Cornec (2010) about these two points. The Central Bank of England fan chart seems more adaptive to the situation but is unfortunately not reproducible as forecasters includes subjective information. In Cornec (2010) a reproducible density forecasting method based on quantile regressions is proposed and gives good results in practice. However, this method did not receive any theoretical support up to our knowledge. The primary motivation of the current paper was to provide a theoretical support to Cornec (2010).

5.8.2 Application of Theorem 5.5.1

We define X_t as the information that becomes available at time t, $X_t = (\Delta\text{GDP}_t, I_t)' \in \mathbb{R}^2$. The loss function will only take into account ΔGDP_t as this is the quantity of interest. We use the quantile loss function (see Example 5.5.2 page 81):

$$\ell_\tau((\Delta\text{GDP}_t, I_t), (\Delta'\text{GDP}_t, I'_t)) = \begin{cases} \tau\left(\Delta\text{GDP}_t - \Delta'\text{GDP}_t\right), & \text{if } \Delta\text{GDP}_t - \Delta'\text{GDP}_t > 0 \\ -\left(1-\tau\right)\left(\Delta\text{GDP}_t - \Delta'\text{GDP}_t\right), & \text{otherwise.} \end{cases}$$

To remind that the risk depends on τ, we add a subscript τ in the notation $R^\tau(\theta) := \mathbb{E}\left[\ell_\tau\left(\Delta\text{GDP}_t, f_\theta(X_{t-1}, X_{t-2})\right)\right]$ and let r_n^τ denote the associated empirical risk. We use the family of predictors proposed by Cornec (2010). The reason is that one of the conclusions of Cornec (2010); Li (2010) is that this set of predictors allow to obtain a forecasting as accurate as the INSEE. It is given by

$$f_\theta(X_{t-1}, X_{t-2}) = \theta_0 + \theta_1 \Delta\text{GDP}_{t-1} + \theta_2 I_{t-1} + \theta_3 (I_{t-1} - I_{t-2})|I_{t-1} - I_{t-2}| \quad (5.4)$$

where $\theta = (\theta_0, \theta_1, \theta_2, \theta_3) \in \Theta(B)$. Fix $R > 0$ and

$$\Theta = \left\{\theta = (\theta_0, \theta_1, \theta_2, \theta_3) \in \mathbb{R}^4, \|\theta\|_1 = \sum_{i=0}^{3} |\theta_i| \leq R\right\}.$$

Remark that in this framework, Assumption **Lip** is satisfied with $L = R + 1$, and the loss function is K-Lipschitz with $K = 1$ so Assumption **LipLoss** is also satisfied. We compare the performance of both ERM and Gibbs estimator.

Corollary 5.8.1. *Let us fix $\tau \in (0,1)$. Let us assume that Assumption* **WeakDep** *is satisfied, and that $n \geq \max\left(10, \kappa^2/(3\mathcal{B}^2)\right)$. Let us fix $\lambda = \sqrt{3n}/\kappa$. Then, with probability at least $1-\varepsilon$ we have*

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_{\theta\in\Theta(B)} \left\{ R^\tau(\theta) + \frac{2\sqrt{3}\kappa}{\sqrt{n}} \left[2.25 + \log\left(\frac{(R+1)\mathcal{B}\sqrt{n}}{\kappa} \right) + \frac{\log\left(\frac{1}{\varepsilon}\right)}{3} \right] \right\}.$$

Remark 5.8.1. *The choice of λ proposed in the theorem may be a problem as in practice we will not know κ. Note that from the proof, it is obvious that in any case, for n large enough, when $\lambda = \sqrt{n}$ we still have a bound*

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_{\theta\in\Theta(B)} \left\{ R^\tau(\theta) + \frac{C(B, \mathcal{B}, \kappa, \varepsilon)}{\sqrt{n}} \right\}.$$

We let $\hat{\theta}^{ERM,\tau}$ denote the ERM with quantile loss ℓ_τ:

$$\hat{\theta}^{ERM,\tau} \in \arg\min_{\theta\in\Theta} r_n^\tau(\theta).$$

We apply Theorem 5.6.6. Note that Assumption **Lip**(L) is satisfied Θ' with $L = R + 1$, Assumption **LipLoss**(K) is satisfied with $K = 1$. Finally, under **WeakDep**$(\mathcal{B}, \mathcal{C})$, the assumptions of Theorem 5.6.6 are satisfied with $\psi = \mathcal{B}$ and $d = 4$.

Corollary 5.8.2. *Let us fix $\tau \in (0,1)$. Let us assume that Assumption* **WeakDep**$(\mathcal{B}, \mathcal{C})$ *is satisfied, Then we have, for any $\varepsilon > 0$ and for n large enough,*

$$\mathbb{P}\left\{ R^\tau(\hat{\theta}^{ERM,\tau}) \leq \inf_{\theta\in\Theta} R^\tau(\theta) + \frac{2\kappa\sqrt{2}}{\sqrt{n}\left(1-\frac{4}{n}\right)} \log\left(\frac{2e^2\mathcal{B}(R+1)\sqrt{n}}{\kappa\varepsilon}\right)\right\} \geq 1-\varepsilon.$$

In the simulations, it appears that the choice of R has little importance as soon as R is large enough: in this case, the simulation shows that the estimator does not really depend on R - only the theoretical bound does. As a consequence we take $R = 100$ in our experiments.

5.8.3 Results

The results are shown in Figure 5.1 for prediction, $\tau = 0.5$, in Figure 5.2 for confidence interval of order 50%, i.e. $\tau = 0.25$ and $\tau = 0.75$ (left) and for confidence interval of order 90%, i.e. $\tau = 0.05$ and $\tau = 0.95$ (right). We report only the results for the period 2000-Q1 to 2011-Q3 (using the period 1988-Q1 to 1999-Q4 for learning).

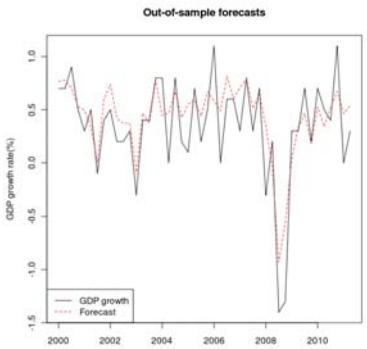

Figure 5.1: French GDP online prediction using the quantile loss function with $\tau = 0.5$.

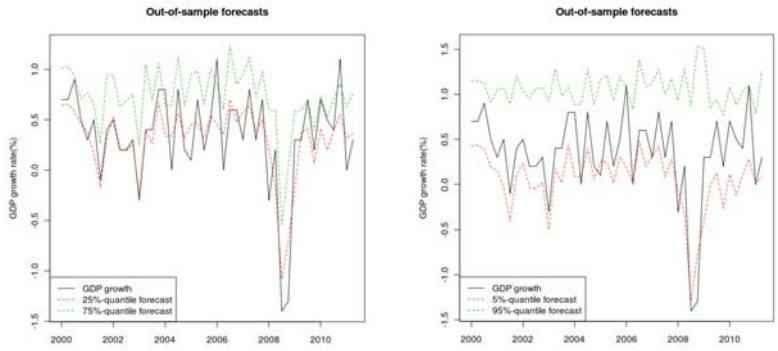

Figure 5.2: French GDP online 50%-confidence intervals (left) and 90%-confidence intervals (right).

We denote $\hat{\theta}^{ERM,\tau}[t']$ the estimator computed at time t', based on the observations from $t = 1$ to $t = t' - 1$. We report the online performance:

$$\begin{aligned} \text{mean abs. pred. error} &= \frac{1}{n}\sum_{t=1}^{n}\left|\Delta GDP_t - f_{\hat{\theta}^{ERM,0.5}[t]}(X_{t-1}, X_{t-2})\right| \\ \text{mean quad. pred. error} &= \frac{1}{n}\sum_{t=1}^{n}\left[\Delta GDP_t - f_{\hat{\theta}^{ERM,0.5}[t]}(X_{t-1}, X_{t-2})\right]^2 \end{aligned}$$

and compare it to the INSEE performance, see Table 5.1.

We also report the frequency of realizations of the GDP falling above the predicted τ-quantile for each τ, see Table 5.2. Note that this quantity should be close to τ.

We completely fail to forecast the 2008 subprime crisis. However, as noted in Cornec (2010), the INSEE forecast for that quarter was also completely wrong. This is in accordance with the fact mentionned above that it is more difficult to forecast the GDP during crisis. However, it is interesting to note that our confidence interval shows that our prediction at this date is less reliable than the previous ones: so, at this time, the forecasters could have been aware that their prediction was unreliable.

One of the most interesting point is to remark that the lower bound of the predicted confidence intervals are really varying over time, while the upper bound is almost constant in the case of $\tau = 0.95$. This is another evidence that the distribution of the errors is non symmetric, and that a parametric model with gaussian innovations would lead to clearly underestimate the magnitude of recessions.

Predictor	Mean absolute prediction error	Mean quadratic prediction error
$\widehat{\theta}^{ERM,0.5}$	0.2249	0.0812
INSEE	0.2579	0.0967

Table 5.1: *Performances of the ERM and of the INSEE.*

τ	Estimator	Frequency
0.05	$\widehat{\theta}^{ERM,0.05}$	0.1739
0.25	$\widehat{\theta}^{ERM,0.25}$	0.4130
0.5	$\widehat{\theta}^{ERM,0.5}$	0.6304
0.75	$\widehat{\theta}^{ERM,0.75}$	0.9130
0.95	$\widehat{\theta}^{ERM,0.95}$	0.9782

Table 5.2: *Empirical frequencies of the event: GDP falls under the predicted τ-quantile.*

5.9 Simulation study

5.9.1 First case: parametric family of predictors

The ERM estimator is now compared to parametric estimators assuming an ARMA form for the time series on a set of simulated data. Here again we consider the ERM estimator for both the quadratic and absolute loss. We compare the performances of both estimators to the one computed by the R procedure "arma" R.

We consider observations drawn from an AR(1) models and a slight variant, see (5.5) and (5.6). Namely, we simulate sequences of length $n = 100$ and $n = 1000$ from the following first-order autoregressive processes:

$$X_t = 0.5X_{t-1} + \varepsilon_t \tag{5.5}$$

$$X_t = 0.5\sin(X_{t-1}) + \varepsilon_t \tag{5.6}$$

where ε_t is the iid innovation. We consider two cases of distributions for ε_t: the uniform case, $\varepsilon_t \sim \mathcal{U}[-a, a]$, and the Gaussian case, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. Note that, in the first case, our two models satisfy the assumptions of Theorem 5.5.1 and Theorem 5.7.1. More precisely there exists a stationary solutions (X_t) that is

n	Model	Innovations	ERM abs.	ERM quad.	ARMA
100	(5.5)	Gaussian	**0.1538**	0.1549	0.1577
		Uniform	**0.1716**	0.1739	0.1774
	(5.6)	Gaussian	0.1714	**0.1705**	0.1736
		Uniform	0.1512	**0.1510**	0.1542
1000	(5.5)	Gaussian	**0.1652**	0.1659	0.1662
		Uniform	**0.1553**	0.1558	0.1562
	(5.6)	Gaussian	0.1545	**0.1526**	0.1530
		Uniform	0.1767	**0.1760**	0.1764

Table 5.3: *Performances of the ERM estimator and ARMA, on the simulations. We highlight the best result for each experiment. The first row "ERM abs." is for the ERM estimator with absolute loss, the second row "ERM quad." for the ERM with quadratic loss.*

ϕ-mixing for an $AR(p)$ process with uniform innovations. and as a consequence $\mathbf{WeakDep}(\mathcal{B}, \mathcal{C})$ is satisfied. In the Gaussian case, however, it is known that $\{X_t\}$ is no longer ϕ-mixing, see Doukhan (1994). However, as this case is more classical in statistics, it is worth testing if our method performs well in practice in this case too.

We take $\sigma = 0.4$ and $a = 0.70$. In both cases this leads to $Var(\epsilon_t) \simeq 0.16$. For each model, we simulate first a sequence of length n, we take the observations 1 to $n-1$ as a learning set and we predict X_n. Each simulation is repeated 100 times and we report the mean error of each method on the Table 5.3. The evolution of the performance is measured by the quadratic prevision error.

It is interesting to note that the ERM estimator with absolute loss performs better on model (5.5) while the ERM with quadratic loss performs slightly better on model (5.6). The differences might be too small to be significative, however, the numerical results tends to indicate that both methods are robust to model mispecification. Also, both estimators seem to perform better than the R "arma" procedure when $n = 100$, but the differences tends to be less perceptible when n grows.

5.9.2 Second case: sparse autoregression

Here, we illustrate Corollary 5.7.1. We compare here the Gibbs estimator to the model selection approach of the "arma" procedure in the R software. This procedure computes the parametric estimator in each submodel AR(p) and then selects the order p by Akaike's AIC criterion Akaike (1973). Note that the computation of the Gibbs estimator in this case is described in Alquier and Lounici (2011) using a Reversible Jump MCMC algorithm. For the parameter λ, $\lambda = n/\hat{\text{var}}(X)$, where $\hat{\text{var}}(X)$ is the empirical variance of the observed time series.

We generate the data according to the following:

$$X_t = 0.5X_{t-1} + 0.1X_{t-2} + \varepsilon_t \tag{5.7}$$

$$X_t = 0.6X_{t-4} + 0.1X_{t-8} + \varepsilon_t \tag{5.8}$$

$$X_t = \cos(X_{t-1})\sin(X_{t-2}) + \varepsilon_t \tag{5.9}$$

where ε_t is the innovation. We still use two models for the innovation: the uniform case, $\varepsilon_t \sim \mathcal{U}[-a, a]$, and the Gaussian case, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. Also we still take $\sigma = 0.4$ and $a = 0.70$. We compare the Gibbs estimator performances to the ones of AIC criterion as implemented in the R software and to the basic least square estimator in the model $AR(q)$ - that we will call "full model". The experimental design is the following: for each model, we simulate a time series of length $2n$, use the observations 1 to n as a learning set and $n+1$ to $2n$ as a test set. We report the performances on the test set. We take $n = 100$ and $n = 1000$ in the simulations. Each simulation is repeated 20 times, we report on Table 5.4 the mean performance and standard deviation of each method.

It is interesting to note that our estimator performs better on Model (5.8) and Model (5.9) while AIC performs slightly better on Model (5.7). The differences tends to be less perceptible when n grows - this is coherent with the fact that we develop here a non-asymptotic theory. It is also interesting to note that our estimator seems to perform well even in the case of a Gaussian noise.

Table 5.4: *Performances of the Gibbs estimator, AIC and least square estimator in the full model, on the simulations. We reported the mean performance and standard deviation of each method. We highlight the best result for each experiment.*

n	Model	Innovations	Gibbs	AIC	Full Model
100	(5.7)	Uniform	**0.165** (0.022)	**0.165** (0.023)	0.182 (0.029)
		Gaussian	0.167 (0.023)	**0.161** (0.023)	0.173 (0.027)
	(5.8)	Uniform	**0.163** (0.020)	0.169 (0.022)	0.178 (0.022)
		Gaussian	**0.172** (0.033)	0.179 (0.040)	0.201 (0.049)
	(5.9)	Uniform	**0.174** (0.022)	0.179 (0.028)	0.201 (0.040)
		Gaussian	**0.179** (0.025)	0.182 (0.025)	0.202 (0.031)
1000	(5.7)	Uniform	**0.163** (0.005)	**0.163** (0.005)	0.166 (0.005)
		Gaussian	**0.160** (0.005)	**0.160** (0.005)	0.162 (0.005)
	(5.8)	Uniform	**0.164** (0.004)	0.166 (0.004)	0.167 (0.004)
		Gaussian	**0.160** (0.008)	0.161 (0.008)	0.163 (0.008)
	(5.9)	Uniform	**0.171** (0.005)	0.172 (0.006)	0.175 (0.006)
		Gaussian	**0.173** (0.009)	**0.173** (0.009)	0.176 (0.010)

5.10 Proofs

5.10.1 Preliminaries

Lemma 5.10.1. *We assume that* **LowRates**(κ) *is satisfied for some* $\kappa > 0$*. For any* $\lambda > 0$ *and* $\theta \in \Theta$ *we have*

$$\mathbb{E}\Big(e^{\lambda(R(\theta)-r_n(\theta))}\Big) \vee E\Big(e^{\lambda(r_n(\theta)-R(\theta))}\Big) \leq \exp\left(\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}\right).$$

Proof of Lemma 5.10.1. Let us fix $\lambda > 0$ and $\theta \in \Theta$. Let us define the function h by:

$$h(x_1, \ldots, x_n) = \frac{1}{K(1+L)} \sum_{i=k+1}^{n} \ell(f_\theta(x_{i-1}, \ldots, x_{i-k}), x_i).$$

We now check that h satisfies (5.1), remember that $\ell(x, x') = g(x - x')$ so

$$\begin{aligned}
&\Big|h\left(x_1, \ldots, x_n\right) - h\left(y_1, \ldots y_n\right)\Big| \\
&\quad \leq \frac{1}{K(1+L)} \sum_{i=k+1}^{n} \Big|g(f_\theta(x_{i-1}, \ldots, x_{i-k}) - x_i) - g(f_\theta(y_{i-1}, \ldots, y_{i-k}) - y_i)\Big| \\
&\quad \leq \frac{1}{1+L} \sum_{i=k+1}^{n} \Big\|\Big(f_\theta(x_{i-1}, \ldots, x_{i-k}) - x_i\Big) - \Big(f_\theta(y_{i-1}, \ldots, y_{i-k}) - y_i\Big)\Big\|
\end{aligned}$$

where we used Assumption **LipLoss**(K) for the last inequality. So we have

$$\Big|h\left(x_1, \ldots, x_n\right) - h\left(y_1, \ldots y_n\right)\Big|$$

$$\leq \frac{1}{1+L}\sum_{i=k+1}^{n}\left(\Big\|f_\theta(x_{i-1},\dots,x_{i-k})-f_\theta(y_{i-1},\dots,y_{i-k})\Big\|+\Big\|x_i-y_i\Big\|\right)$$
$$\leq \frac{1}{1+L}\sum_{i=k+1}^{n}\left(\sum_{j=1}^{k}a_j(\theta)\|x_{i-j}-y_{i-j}\|+\|x_i-y_i\|\right)$$
$$\leq \frac{1}{1+L}\sum_{i=1}^{n}\left(1+\sum_{j=1}^{k}a_j(\theta)\right)\|x_i-y_i\| \leq \sum_{i=1}^{n}\|x_i-y_i\|$$

where we used Assumption $\mathbf{Lip}(L)$. So we can apply Lemma 5.3.2 with $h(X_1,\dots,X_n) = \frac{n-k}{K(1+L)}r_n(\theta)$, $\mathbb{E}(h(X_1,\dots,X_n)) = \frac{n-k}{K(1+L)}R(\theta)$, and $t = K(1+L)\lambda/(n-k)$:

$$\mathbb{E}\left(e^{\lambda[R(\theta)-r_n(\theta)]}\right) \leq \exp\left(\frac{\lambda^2K^2(1+L)^2\left(\mathcal{B}+\theta_{\infty,n}(1)\right)^2}{2n\left(1-k/n\right)^2}\right)$$
$$\leq \exp\left(\frac{\lambda^2K^2(1+L)^2\left(\mathcal{B}+\mathcal{C}\right)^2}{2n\left(1-\frac{k}{n}\right)^2}\right)$$

by Assumption $\mathbf{WeakDep}(\mathcal{C})$. This ends the proof of the first inequality. The reverse inequality is obtained by replacing the function h by $-h$. ■

We are now ready to state the following key Lemma.

Lemma 5.10.2. *Let us assume that* $\mathbf{LowRates}(\kappa)$ *is satisfied satisfied for some* $\kappa > 0$. *Then for any* $\lambda > 0$ *we have*

$$\mathbb{P}\left\{\begin{array}{l}\forall\rho\in\mathcal{M}_+^1(\Theta),\\ \int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\\ \textit{and}\\ \int r_n\mathrm{d}\rho \leq \int R\mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\end{array}\right\} \geq 1-\varepsilon. \tag{5.10}$$

Proof of Lemma 5.10.2. Let us fix $\theta > 0$ and $\lambda > 0$, and apply the first inequality of Lemma 5.10.1. We have:

$$\mathbb{E}\Big(\exp\left(\lambda\Big(R(\theta)-r_n(\theta)-\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}\Big)\right)\Big) \leq 1,$$

and we multiply this result by $\varepsilon/2$ and integrate it with respect to $\pi(\mathrm{d}\theta)$. An application of Fubini's Theorem yields

$$\mathbb{E}\int\exp\left(\lambda(R(\theta)-r_n(\theta))-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)\right)\pi(\mathrm{d}\theta) \leq \frac{\varepsilon}{2}.$$

We apply Lemma 5.3.1 and we get:

$$\mathbb{E}\exp\left(\sup_{\rho}\left\{\lambda\int(R(\theta)-r_n(\theta))\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)-\mathcal{K}(\rho,\pi)\right\}\right)\leq\frac{\varepsilon}{2}.$$

As $e^x\geq\mathbf{1}_{\mathbb{R}_+}(x)$, we have:

$$\mathbb{P}\left\{\sup_{\rho}\left\{\lambda\int\left(R(\theta)-r_n(\theta)\right)\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

Using the same arguments than above but starting with the second inequality of Lemma 5.10.1:

$$\mathbb{E}\exp\left(\lambda\Big(r_n(\theta)-R(\theta)-\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}\Big)\right)\leq 1.$$

we obtain:

$$\mathbb{P}\left\{\sup_{\rho}\left\{\lambda\int\left[r_n(\theta)-R(\theta)\right]\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}-\log\left(\frac{2}{\varepsilon}\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

A union bound ends the proof. ■

The following variant of Lemma 5.10.2 will also be useful.

Lemma 5.10.3. *Let us assume that* **LowRates**(κ) *is satisfied satisfied for some* $\kappa>0$. *Then for any* $\lambda>0$ *we have*

$$\mathbb{P}\left\{\begin{array}{l}\forall\rho\in\mathcal{M}_+^1(\Theta),\\ \int R\mathrm{d}\rho\leq\int r_n\mathrm{d}\rho+\frac{\lambda\kappa^2}{n(1-k/n)^2}+\frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\\ \quad and\\ r_n(\overline{\theta})\leq R(\overline{\theta})+\frac{\lambda\kappa^2}{n(1-k/n)^2}+\frac{\log(2/\varepsilon)}{\lambda}\end{array}\right\}\geq 1-\varepsilon.$$

Proof of Lemma 5.10.3. Following the proof of Lemma 5.10.2 we have:

$$\mathbb{P}\left\{\sup_{\rho}\left\{\lambda\int\left(R(\theta)-r_n(\theta)\right)\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

Now, we use the second inequality of Lemma 5.10.1, with $\theta=\overline{\theta}$:

$$\mathbb{E}\Big(\exp\left(\lambda\Big(r_n(\overline{\theta})-R(\overline{\theta})-\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}\Big)\right)\Big)\leq 1.$$

But then, we directly apply Markov's inequality to get:

$$\mathbb{P}\left\{r_n(\overline{\theta})\geq R(\overline{\theta})+\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{\log\left(2/\varepsilon\right)}{\lambda}\right\}\leq\frac{\varepsilon}{2}.$$

Here again, a union bound ends the proof. ■

5.10.2 Proof of Theorems 5.5.1 , 5.6.5 and 5.6.7

Proof of Theorem 5.5.1. We apply Lemma 5.10.2. So, with probability at least $1-\varepsilon$ we are on the event given by (5.10). From now, we work on that event. The first inequality of (5.10), when applied to $\hat{\rho}_\lambda(\mathrm{d}\theta)$, gives

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \int r_n(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{1}{\lambda}\log\left(2/\varepsilon\right) + \frac{1}{\lambda}\mathcal{K}(\hat{\rho}_\lambda,\pi).$$

According to Lemma 5.3.1 we have:

$$\int r_n(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\hat{\rho}_\lambda,\pi) = \inf_\rho\left(\int r_n(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right)$$

so we obtain

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\rho\left\{\int r_n(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi)+\log\left(2/\varepsilon\right)}{\lambda}\right\}. \tag{5.11}$$

We now estimate from above $r(\theta)$ by $R(\theta)$. Applying the second inequality of (5.10) and plugging it into Inequality 5.11 gives

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\rho\left\{\int R\mathrm{d}\rho + \frac{2}{\lambda}\mathcal{K}(\rho,\pi) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2}{\lambda}\log\left(2/\varepsilon\right)\right\}.$$

We end the proof by the remark that $\theta \mapsto R(\theta)$ is convex and so by Jensen's inequality $\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \geq R\left(\int\theta\hat{\rho}_\lambda(\mathrm{d}\theta)\right) = R(\hat{\theta}_\lambda)$. ■

Proof of Theorem 5.6.5. An application of Theorem 5.5.1 yields that with probability at least $1-\varepsilon$

$$R(\hat{\theta}_\lambda) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)}\left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\mathcal{K}(\rho,\pi)+2\log\left(2/\varepsilon\right)}{\lambda}\right].$$

Let us estimate the upper bound at the probability distribution ρ_δ defined as

$$\frac{\mathrm{d}\rho_\delta}{\mathrm{d}\pi}(\theta) = \frac{\mathbf{1}\{R(\theta)-R(\overline{\theta})<\delta\}}{\int_{t\in\Theta}\mathbf{1}\{R(t)-R(\overline{\theta})<\delta\}\pi(\mathrm{d}t)}.$$

Then we have:

$$R\left(\hat{\theta}_\lambda\right) \leq \inf_{\delta>0}\Bigg[R(\overline{\theta}) + \delta + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + 2\frac{-\log\int_{t\in\Theta}\mathbf{1}\{R(t)-\inf_\Theta R<\delta\}\pi(\mathrm{d}t) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}\Bigg].$$

Under the assumptions of Theorem 5.6.5 we have:

$$R\left(\hat{\theta}_\lambda\right) \leq \inf_{\delta>0}\left[R(\overline{\theta}) + \delta + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + 2\frac{d\log\left(D/\delta\right) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right].$$

The infimum is reached for $\delta = d/\lambda$ and we have:

$$R\left(\hat{\theta}_\lambda\right) \leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + 2\frac{d\log\left(D\sqrt{e}\lambda/d\right) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}.$$

■

Proof of Theorem 5.6.7. Let us apply Lemma 5.10.2 in each model Θ_j, with a fixed $\lambda_j > 0$ and confidence level $\varepsilon_j > 0$. We obtain, for all j,

$$\mathbb{P}\left\{\begin{array}{l} \forall\rho\in\mathcal{M}_+^1(\Theta_j), \\ \int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log(2/\varepsilon_j)}{\lambda_j} \\ \text{ and} \\ \int r_n\mathrm{d}\rho \leq \int R\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log(2/\varepsilon_j)}{\lambda_j} \end{array}\right\} \geq 1-\varepsilon_j.$$

We put $\varepsilon_j = p_j\varepsilon$, a union bound gives leads to:

$$\mathbb{P}\left\{\begin{array}{l} \forall j\in\{1,...,M\}, \quad \forall\rho\in\mathcal{M}_+^1(\Theta_j), \\ \int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j} \\ \text{ and} \\ \int r_n\mathrm{d}\rho \leq \int R\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j} \end{array}\right\} \geq 1-\varepsilon. \tag{5.12}$$

From now, we only work on that event of probability at least $1-\varepsilon$. Remark that

$$\begin{aligned} R(\hat{\theta}) &= R(\hat{\theta}_{\lambda_{\hat{j}},\hat{j}}) \\ &\leq \int R(\theta)\hat{\rho}_{\lambda_{\hat{j}},\hat{j}}(\mathrm{d}\theta) \text{ by Jensen's inequality} \\ &\leq \int r_n\hat{\rho}_{\lambda_{\hat{j}},\hat{j}}(\mathrm{d}\theta) + \frac{\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\hat{\rho}_{\lambda_{\hat{j}},\hat{j}},\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j} \\ &\qquad \text{by (5.12)} \\ &= \inf_{1\leq j\leq M}\left\{\int r_n\hat{\rho}_{\lambda_j,j}(\mathrm{d}\theta) + \frac{\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\hat{\rho}_{\lambda_j,j},\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\right\} \\ &\qquad \text{by definition of } \hat{j} \\ &= \inf_{1\leq j\leq M}\inf_{\rho\in\mathcal{M}_+^1(\Theta_j)}\left\{\int r_n\rho(\mathrm{d}\theta) + \frac{\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\right\} \end{aligned}$$

by Lemma 5.3.1

$$\leq \inf_{1\leq j\leq M} \inf_{\rho\in\mathcal{M}_+^1(\Theta_j)} \left\{ \int R\rho(\mathrm{d}\theta) + \frac{2\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{\mathcal{K}(\rho,\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j} \right\}$$

by (5.12) again

$$\leq \inf_{1\leq j\leq M} \inf_{\delta>0} \left\{ R(\overline{\theta}_j) + \delta + \frac{2\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j\log\left(D_j/\delta\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j} \right\}$$

by restricting ρ as in the proof of Cor. 5.6.5 page 87

$$\leq \inf_{1\leq j\leq M} \left\{ R(\overline{\theta}_j) + \frac{2\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j\log\left(\frac{D_j e\lambda_j}{d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j} \right\}$$

by taking $\delta = \dfrac{d_j}{\lambda_j}$

$$= \inf_{1\leq j\leq M} \left\{ R(\overline{\theta}_j) + \inf_{\lambda>0} \left\{ \frac{2\lambda\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j\log\left(\frac{D_j e\lambda}{d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\} \right\}$$

by definition of λ_j

$$\leq \inf_{1\leq j\leq M} \left\{ R(\overline{\theta}_j) + 2\frac{\kappa_j}{1-k/n} \left\{ \sqrt{\frac{d_j}{n}} \log\left(\frac{D_j e^2}{\kappa_j}\sqrt{\frac{n}{d_j}}\right) + \frac{\log\left(\frac{2}{\varepsilon p_j}\right)}{\sqrt{nd_j}} \right\} \right\}.$$

■

5.10.3 Proof of Theorems 5.6.2 and 5.6.6

Let us now prove the results about the ERM.

Proof of Theorem 5.6.2. We choose π as the uniform probability distribution on Θ and $\lambda > 0$. We apply Lemma 5.10.3. So we have, with probability at least $1-\varepsilon$,

$$\begin{cases} \forall \rho \in \mathcal{M}_+^1(\Theta'), & \int R\mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda} \\ \text{and} & r_n(\overline{\theta}) \leq R(\overline{\theta}) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log(2/\varepsilon)}{\lambda}. \end{cases}$$

We restrict the inf in the first inequality to Dirac masses $\rho \in \{\delta_\theta, \theta\in\Theta\}$ and we obtain:

$$\begin{cases} \forall \theta \in \Theta, & R(\theta) \leq r_n(\theta) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log\left(\frac{2M}{\varepsilon}\right)}{\lambda} \\ \text{and} & r_n(\overline{\theta}) \leq R(\overline{\theta}) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log(2/\varepsilon)}{\lambda}. \end{cases}$$

In particular, we apply the first inequality to $\hat{\theta}^{ERM}$. We remind that $\overline{\theta}$ minimizes

R on Θ and that $\hat{\theta}^{ERM}$ minimizes r_n on Θ, and so we have

$$\begin{aligned} R(\hat{\theta}^{ERM}) &\leq r_n(\hat{\theta}^{ERM}) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\log(M) + \log\left(2/\varepsilon\right)}{\lambda} \\ &\leq r_n(\overline{\theta}) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\log(M) + \log\left(2/\varepsilon\right)}{\lambda} \\ &\leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\log(M) + 2\log\left(2/\varepsilon\right)}{\lambda} \\ &\leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda}. \end{aligned}$$

The result still holds if we choose λ as a minimizer of

$$\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda}.$$

■

Proof of Theorem 5.6.6. We put $\Theta' = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D+1\}$. We choose π as the uniform probability distribution on Θ'. We apply Lemma 5.10.3. So we have, with probability at least $1-\varepsilon$,

$$\begin{cases} \forall \rho \in \mathcal{M}_+^1(\Theta'), & \int R\mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda} \\ \text{and} & r_n(\overline{\theta}) \leq R(\overline{\theta}) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log(2/\varepsilon)}{\lambda}. \end{cases}$$

So for any ρ,

$$\begin{aligned} R(\hat{\theta}^{ERM}) &= \int [R(\hat{\theta}^{ERM}) - R(\theta)]\rho(\mathrm{d}\theta) + \int R\mathrm{d}\rho \\ &\leq \int [R(\hat{\theta}^{ERM}) - R(\theta)]\rho(\mathrm{d}\theta) + \int r_n \mathrm{d}\rho + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi) + \log\left(2/\varepsilon\right)}{\lambda} \\ &\leq \int [R(\hat{\theta}^{ERM}) - R(\theta)]\rho(\mathrm{d}\theta) + \int [r_n(\theta) - r_n(\hat{\theta}^{ERM})]\rho(\mathrm{d}\theta) + r_n(\hat{\theta}^{ERM}) \\ &\qquad + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi) + \log\left(2/\varepsilon\right)}{\lambda} \\ &\leq 2K\psi \int \|\theta - \hat{\theta}^{ERM}\|_1 \rho(\mathrm{d}\theta) + r_n(\overline{\theta}) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi) + \log\left(2/\varepsilon\right)}{\lambda} \\ &\leq 2K\psi \int \|\theta - \hat{\theta}^{ERM}\|_1 \rho(\mathrm{d}\theta) + R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi) + 2\log\left(2/\varepsilon\right)}{\lambda}. \end{aligned}$$

Now we define, for any $\delta > 0$, ρ_δ by

$$\frac{\mathrm{d}\rho_\delta}{\mathrm{d}\pi}(\theta) = \frac{\mathbf{1}\{\|\theta - \hat{\theta}^{ERM}\| < \delta\}}{\int_{t\in\Theta'} \mathbf{1}\{\|t - \hat{\theta}^{ERM}\| < \delta\}\pi(\mathrm{d}t)}.$$

So in particular, we have, for any $\delta > 0$,

$$R(\hat{\theta}^{ERM}) \leq 2K\psi\delta + R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\log \frac{1}{\int_{t\in\Theta'} \mathbf{1}\{\|t-\hat{\theta}^{ERM}\|<\delta\}\pi(\mathrm{d}t)} + 2\log\left(2/\varepsilon\right)}{\lambda}.$$

But for any $\delta \leq 1$,

$$-\log \int_{t\in\Theta'} \mathbf{1}\{\|t-\hat{\theta}^{ERM}\| < \delta\}\pi(\mathrm{d}t) = d\log\left(\frac{D+1}{\delta}\right).$$

So we have

$$R(\hat{\theta}^{ERM}) \leq \inf_{\delta\leq 1}\left\{2K\psi\delta + R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{d\log\left(\frac{D+1}{\delta}\right) + 2\log\left(2/\varepsilon\right)}{\lambda}\right\}.$$

We optimize this result by taking $\delta = d/(2\lambda K\psi)$, which is smaller than 1 as soon as $t \geq 2K\psi/d$, we get:

$$R(\hat{\theta}^{ERM}) \leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{d\log\left(\frac{2eK\psi(D+1)t}{d}\right) + 2\log\left(2/\varepsilon\right)}{\lambda}.$$

We just choose λ as the minimizer of the r.h.s., subject to $t \geq 2K\psi/d$, to end the proof. ■

5.10.4 Some preliminary lemmas for the proof of Theorem 5.7.1

Lemma 5.10.4. *Under the hypothesis of Theorem 5.7.1, we have, for any $\theta \in \Theta$, for any $0 \leq \lambda \leq (n-k)/(2kKL\mathcal{BC})$,*

$$\mathbb{E}\exp\left\{\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(R(\theta)-R(\overline{\theta})\right) - r(\theta) + r(\overline{\theta})\right]\right\} \leq 1,$$

and

$$\mathbb{E}\exp\left\{\lambda\left[\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(R(\overline{\theta})-R(\theta)\right) - r(\overline{\theta}) + r(\theta)\right]\right\} \leq 1.$$

Lemma 5.10.4. We apply Lemma 5.3.3 to $N = n-k$, $Z_i = (X_{i+1}, \ldots, X_{i+k})$,

$$f(Z_i) = \frac{1}{n-k}\Big[R(\theta) - R(\overline{\theta}) - \ell\left(X_{i+k}, f_\theta(X_{i+k-1},\ldots,X_{i+1})\right) + \ell\left(X_{i+k}, f_{\overline{\theta}}(X_{i+k-1},\ldots,X_{i+1})\right)^2\Big],$$

and so

$$S_N(f) = [R(\theta) - R(\overline{\theta}) - r(\theta) + r(\overline{\theta})],$$

and the Z_i are uniformly mixing with coefficients $\phi_r^Z = \phi_{\lfloor r/q \rfloor}$. Note that $1 + \sum_{r=1}^{n-q} \sqrt{\phi_r^Z} = 1 + \sum_{r=1}^{n-q} \sqrt{\phi_{\lfloor r/k \rfloor}} \leq k\,\mathcal{C}$ by **PhiMix**($\mathcal{C}$). For any θ and θ' in Θ let us put

$$V(\theta, \theta') = \mathbb{E}\left\{ \left[\ell\Big(X_{k+1}, f_\theta(X_k, ..., X_1)\Big) - \ell\Big(X_{k+1}, f_{\theta'}(X_k, ..., X_1)\Big) \right]^2 \right\}.$$

We are going to apply Lemma 5.3.3. Remark that $\sigma^2(f) \leq V(\theta, \overline{\theta})/(n-k)^2$. Also,

$$\begin{aligned} &\Big| \ell\left(X_{i+k}, f_\theta(X_{i+k-1}, \ldots, X_{i+1})\right) - \ell\left(X_{i+k}, f_{\overline{\theta}}(X_{i+k-1}, \ldots, X_{i+1})\right) \Big| \\ &\qquad \leq K \left| f_\theta(X_{i+k-1}, \ldots, X_{i+1}) - f_{\overline{\theta}}(X_{i+k-1}, \ldots, X_{i+1}) \right| \leq KL\mathcal{B} \end{aligned}$$

where we used LipLoss(K) for the first inequality and Lip(L) and PhiMix($\mathcal{B}, \mathcal{C}$) for the second inequality. This implies that $\|f\|_\infty \leq 2KL\mathcal{B}/(n-k)$, so we can apply Lemma 5.3.3 for any $0 \leq \lambda \leq (n-k)/(2kKL\mathcal{B}\mathcal{C})]$, we have

$$\ln \mathbb{E} \exp\left[\lambda \Big(R(\theta) - R(\overline{\theta}) - r(\theta) + r(\overline{\theta}) \Big) \right] \leq \frac{8k\mathcal{C}V(\theta, \overline{\theta})\lambda^2}{n-k}.$$

Notice finally that Margin($\mathcal{K}$) leads to

$$V(\theta, \overline{\theta}) = \mathcal{K}\left[R(\theta) - R(\overline{\theta}) \right]$$

This proves the first inequality of Lemma 5.10.4. The second inequality is proved exacly in the same way, but replacing f by $-f$. □

We are now ready to state the following key Lemma.

Lemma 5.10.5. *Under the hypothesis of Theorem 5.7.1, we have, for any* $0 \leq \lambda \leq (n-k)/(2kKL\mathcal{B}\mathcal{C})$*, for any* $0 < \varepsilon < 1$*,*

$$\mathbb{P}\left\{ \begin{array}{l} \forall \rho \in \mathcal{M}_+^1(\Theta), \\ \left(1 - \frac{8k\mathcal{C}\lambda}{n-k}\right)\left(\int R \mathrm{d}\rho - R(\overline{\theta})\right) \leq \int r \mathrm{d}\rho - r(\overline{\theta}) + \frac{\mathcal{K}(\rho,\pi) + \log(2/\varepsilon)}{\lambda} \\ \quad \textit{and} \\ \int r \mathrm{d}\rho - r(\overline{\theta}) \leq \left(\int R \mathrm{d}\rho - R(\overline{\theta})\right)\left(1 + \frac{8k\mathcal{C}\lambda}{n-k}\right) + \frac{\mathcal{K}(\rho,\pi) + \log(2/\varepsilon)}{\lambda} \end{array} \right\} \geq 1 - \varepsilon.$$

Proof of Lemma 5.10.5. Let us fix ε, λ and $\theta \in \Theta$, and apply the first inequality of Lemma 5.10.4. We have:

$$\mathbb{E}\exp\left\{\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\Big(R(\theta)-R(\overline{\theta})\Big)-r(\theta)+r(\overline{\theta})\right]\right\}\leq 1,$$

and we multiply this result by $\varepsilon/2$ and integrate it with respect to $\pi(\mathrm{d}\theta)$. Fubini's Theorem gives:

$$\mathbb{E}\int\exp\left\{\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\Big(R(\theta)-R(\overline{\theta})\Big)-r(\theta)+r(\overline{\theta})+\log(\epsilon/2)\right]\right\}\pi(\mathrm{d}\theta) \leq \frac{\varepsilon}{2}.$$

We apply Lemma 5.3.1 and we get:

$$\mathbb{E}\exp\left\{\sup_{\rho}\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\Big(\int R\mathrm{d}\rho-R(\overline{\theta})\Big)-\int r\mathrm{d}\rho+r(\overline{\theta}) +\log(\epsilon/2)-\mathcal{K}(\rho,\pi)\right]\right\}\leq\frac{\varepsilon}{2}.$$

As $e^x \geq \mathbf{1}_{\mathbb{R}_+}(x)$, we have:

$$\mathbb{P}\left\{\sup_{\rho}\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\Big(\int R\mathrm{d}\rho-R(\overline{\theta})\Big)-\int r\mathrm{d}\rho+r(\overline{\theta}) +\log(\epsilon/2)\right]-\mathcal{K}(\rho,\pi)\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

Let us apply the same arguments starting with the second inequality of Lemma 5.10.4. We obtain:

$$\mathbb{P}\left\{\sup_{\rho}\lambda\left[\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\Big(R(\overline{\theta})-\int R\mathrm{d}\rho\Big)-r(\overline{\theta})+\int r\mathrm{d}\rho +\log(\epsilon/2)-\mathcal{K}(\rho,\pi)\right]\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

A union bound ends the proof. ∎

5.10.5 Proof of Theorem 5.7.1

Proof of Theorem 5.7.1. Fix $0 \leq \lambda = (n-k)/(4kKL\mathcal{BC}) \wedge (n-k)/(16k\mathcal{C}) \leq (n-k)/(2kKL\mathcal{BC})$. Applying Lemma 5.10.5, we assume from now that the event

of probability at least $1-\varepsilon$ given by this lemma is satisfied. In particular we have $\forall \rho \in \mathcal{M}_+^1(\Theta)$,

$$\int R \mathrm{d}\rho - R(\overline{\theta}) \leq \frac{\int r \mathrm{d}\rho - r(\overline{\theta}) + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)}.$$

In particular, thanks to Lemma 5.3.1, we have:

$$\int R \mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \frac{\int r \mathrm{d}\rho - r(\overline{\theta}) + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)}.$$

Now, we apply the second inequality of Lemma 5.10.5:

$$\begin{aligned}
&\int R \mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) \\
&\qquad \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \frac{\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left[\int R \mathrm{d}\rho - R(\overline{\theta})\right] + 2\frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)} \\
&\qquad \leq \inf_j \inf_{\rho \in \mathcal{M}_+^1(\Theta_j)} \frac{\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left[\int R \mathrm{d}\rho - R(\overline{\theta})\right] + 2\frac{\mathcal{K}(\rho_j,\pi)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)} \\
&\qquad\qquad \leq \inf_j \inf_{\delta>0} \frac{\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left[R(\overline{\theta}_j)+\delta - R(\overline{\theta})\right] + 2\frac{d_j \log\left(\frac{D_j}{\delta}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)}
\end{aligned}$$

by restricting ρ as in the proof of Theorem 5.6.5. First, notice that our choice $\lambda \leq (n-k)/(16k\mathcal{C})$ leads to

$$\begin{aligned}
\int R \mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) &\leq 2 \inf_j \inf_{\delta>0} \left\{ \frac{3}{2}\left[R(\overline{\theta}_j)+\delta - R(\overline{\theta})\right] + 2\frac{d_j \log\left(\frac{D_j}{\delta}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\} \\
&\leq 4 \inf_j \inf_{\delta>0} \left\{ R(\overline{\theta}_j)+\delta - R(\overline{\theta}) + \frac{d_j \log\left(\frac{D_j}{\delta}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\}.
\end{aligned}$$

Taking $\delta = d_j/\lambda$ leads to

$$\int R \mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) \leq 4 \inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + \frac{d_j \log\left(\frac{D_j e \lambda}{d_j}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\}.$$

Finally, we replace the last occurences of λ by its value:

$$\begin{aligned}
&\int R \mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) \\
&\leq 4 \inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + (16k\mathcal{C} \vee 4kKL\mathcal{B}\mathcal{C}) \frac{d_j \log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{n-k} \right\}.
\end{aligned}$$

Jensen's inequality leads to:

$$R\left(\hat{\theta}_\lambda\right) - R(\overline{\theta}) \leq 4 \inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + 4k\mathcal{C}\left(4 \vee KL\mathcal{B}\right) \frac{d_j \log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{n-k} \right\}.$$

■

Bibliography

Agarwal, A. and Duchi, J. C. (2011). The generalization ability of online algorithms for dependent data. *IEEE Trans. Inform. Theory 59(2011) no.1*, 573–587.

Agarwal, A. and Duchi, J. C. and Johansson, M. and Jordan, M. I. (2011). Ergodic Mirror Descent. available at *Preprint arXiv:1105.4681.*

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory, Budapest: Akademia Kiado,* 267-281.

Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) Process. *J. Time Series Anal. 8*(3), 261–275.

Al-Osh, M. A. and Alzaid, A. A. (1990). An Integer-Valued pth-order Autoregressive Structure (INAR(p)) Process. *J. Appl. Prob 27*, 314–324.

Alquier, P. (2008). PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics, 17*, 279-304.

Alquier, P. and Li, X. (2012). Prediction of quantiles by statistical learning and application to GDP forecasting. *in the proceedings of DS'12 (conference on Discovery Science), Springer, Lecture Notes in Artificial Intelligence, 2012*, 22-36.

Alquier, P. and Lounici, P. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics, 5*, 127-145.

Alquier, P. and Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli, 18*, 883-193.

Athreya, K. B. and Pantula, S. G. (1986). Mixing properties of Harris chains and autoregressive processes. *J. Appl. Probab. 23*, 880–892.

Andrews, D. (1984). Non strong mixing autoregressive processes. *J. Appl. Prob 21*, 930–934.

Audibert, J.-Y. (2007). Fast rates in statistical inference through aggregation., *Annals of Statistics 35*, 1591-1646.

Audibert, J.-Y. (2010). PAC-Bayesian aggregation and multi-armed bandits., *HDR Université Paris Est.*

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression., *The Annals of Statistics 35 no. 5*, 2766-2794.

Azoury, K. S. and Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions., *Machine Learning 43 no. 3*, 211-246.

Baraud, Yannick and Comte, F. and Viennet, G. (2001). Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist. 5*, 33-49.

Bardet,J-M. Doukhan, P. and Léon, JR. (2007). *Uniform limit theorems for the integrated periodogram of weakly dependent time series and their applications to WhittleÕs estimate. J.T.S.A. (to appear).*

Belloni, A. and Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist., 39 no. 1*, 82-130.

Berkowitz, J. (2001). Testing Density Forecasts, with Applications to Risk Management. *Journal of Business and Economic Statistics, 19*, 465-474.

Biau, G. and Biau, O. and Rouvière, L. (2008). Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data. *Journal of Business Cycle Measurement and Analysis, 3* , 317-332.

Biau, G. and Patra, B. (2011). Sequential quantile prediction of time series. *IEEE Transactions on Information Theory, 57*, 1664-1674.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist. 1*, 1071Ð1095.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society 3 no. 3*, 203-268.

Bosq, D. (1996). Nonparametric Statistics for Stochastic Processes, Estimation and Prediction.. *Lecture Notes in Statistics. 110*. Springer, New York.

Bradley, R. C. (2007). Introduction to strong mixing conditions. Vol. 1,2 & 3. *Kendrick Press, Heber City, UT*.

Britton, E. and Fisher, P. and Whitley, J. (1998). The Inflation Report Projections: Understanding the Fan Chart. *Bank of England Quarterly Bulletin. 38, no. 1*, 30-37.

Brockwell, P. and Davis, R. (2009). Time Series: Theory and Methods (2nd Edition)3. *Springer*.

Bougerol, P. and Picard. N.(1992) Strict Stationarity of Generalized Autoregressive Processes. *Ann. Prob. 20*, 1714–1730.

Bunea, F. and Tsybakov, A. B. and Wegkamp, M. H.(2007) Aggregation for Gaussian regression. *Annals of Statistics. 35*, 1674-1697.

Bühlmann, P. and van de Geer, S.(2011) Statistics for High-Dimensional Data. *Annals of Statistics, Springer 2011*.

Calistri, E. , Livi, R. and Buiatti, M. (2011). Evolutionary trends of GC/AT distribution patterns in promoters. *Molecular Phylogenetics and Evolution. 60(2)228-35*.

Casella, G. and Robert, C. (2004). Monte Carlo Statistical Methods. *Springer-Verlag*.

Catoni, O. (2007). PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning). *Lecture Notes-Monograph Series, IMS. 56*.

Catoni, O. (2004). Statistical Learning Theory and Stochastic Optimization. *Springer Lecture Notes in Mathematics*.

Catoni, O. (2003). A PAC-Bayesian approach to adaptative classification. *Preprint Laboratoire de Probabilités et Modèles Aléatoires. 2003*.

Catoni, O. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP, to appear.*

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

Clements, M. P. (2004). Evaluating the Bank of England Density Forecasts of Inflation. *Economic Journal. 114*, 844-866.

Clavel, L. and Minodier, C. (2009). A monthly indicator of the french business climate. *Documents de Travail de la DESE.*

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association. 74*, 829–836.

Corder, G.W. and Foreman, D.I. (2009). Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach. *Wiley.*

Cornec, M. (2010). Constructing a conditional GDP fan chart with an application to French business survey data. *30th CIRET Conference, New York.*

Coupier, Y. and Doukhan, P. and Ycart, B. (2006). 0-1 laws for dependent images. *ALEA Lat. Am. J. Probab. Math. Stat. 2*, 157–175.

Cuong, N. V. and Tung Ho, L. S. and Dinh, V. (2013). Generalization and Robustness of Batched Weighted Average Algorithm with V-Geometrically Ergodic Markov Data. *Proceedings of ALT'13 Springer, 2013*, 264-278.

Dalalyan, A. and Salmon, J. (2013). Sharp Oracle Inequalities for Aggregation of Affine Estimators. *Machine Learning, 72*, 39-61.

Dalalyan, A. and Tsybakov, A. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *The Annals of Statistics, 40 no. 4*, 2327Ð2355.

Dedecker, J. and Doukhan, P. and Lang, G. and León, J. R. and Louhichi, S and Prieur, C. (2007). *Weak dependence: With Examples and Applications.* Springer-Verlag, New York: Lecture Notes in Statistics **190**.

Dedecker, J. and Doukhan, P. (2003). A new covariance inequality and applications. *Stochastic Processes and Their Applications. 106*, 63–80.

Dedecker, J., Doukhan, P. and Merlevède, F. (2012). Rates of convergence in the strong invariance principle under projective criteria. *Electron. J. Probab 17*, no16, 1–31.

Dedecker, J. and Prieur, C. (2004). Coupling for $\tau-$Dependent Sequences and Applications, *J. Theo. Prob. 17*, 861–885.

Devilliers, M. (2004). Les enquêtes de conjoncture, *Archives et Documents, INSEE. 101.*

Diebold, F. X. and Tay, A. S. and Wallis, K. F. (1997). Evaluating density forecasts of inflation: the Survey of Professional Forecasters, *Discussion Paper No.48, ESRC Macroeconomic Modelling Bureau, University of Warwick and Working Paper No.6228, National Bureau of Economic Research, Cambridge, Mass.*

Donsker, M. D. and Varadhan, S. S. (1976). Asymptotic evaluation of certain Markov process expectations for large time. III., *Communications on Pure and Applied Mathematics 28*, 389-461.

Doukhan, P. (1994). Mixing: properties and examples., *Lecture Notes in Statistics 85. Springer-Verlag.*

Doukhan, P. and Fokianos, K. and Tjøstheim, D. (2012). On weak dependence conditions for Poisson autoregressions., *Statistics & Probability Letters 82*, 942–948.

Doukhan, P. and G. Lang, (2002). Rates in the empirical central limit theorem for stationary weakly dependent random fields., *Stat. Inference Stoch. Process. 5*, 199-228.

Doukhan, P. and Latour, A. and Oraichi, D. (2006). Simple integer-valued bilinear time series model., *Adv. Appl. Prob. 38*, 559–578.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities., *Stoch. Proc. Appl. 84*, 313–342.

Doukhan, P. and Mayo, N. and Truquet, L. (2009). Weak dependence, models and some applications., *Metrika. 69*(2-3), 199–225.

Doukhan, P. and Prohl, S. and Robert, C. Y.(2011). Subsampling weakly dependent time series and application to extremes (with discussion)., *Test 20*, 447–479.

Doukhan, P. and Teyssière, G. and Winant, P. (2006). A LARCH(∞) vector valued process., *In: Bertail P, Doukhan P, Soulier P (eds) Lecture Note in Statistics, Dependence in Probability and statistics, 187*, 245-258.

Doukhan, P. and Wintenberger, O. (2007). An invariance principle for weakly dependent stationary general models. *Prob. Math. Stat. 27*, 45-73.

Doukhan, P. and Wintenberger, O. (2008). Weakly dependent chains with infinite memory. *Stoch. Proc. Appl. 118*, 1997–2013.

Dowd, K. (2004). The inflation fan charts: An evaluation. *Greek Economic Review*, *23*, 99–111.

Drost, F. C., Akker, R. van den and Werker, B. J. (2008). Note on integer-valued bilinear time series models. *Stat. Probab. Lett. 78*, 992–996.

Du, J.-G. and Li,Y. (1991). The integer valued autoregressive (INAR(p)) model. *Time Series Anal. 12*, 129–142.

Dubois, E. and Michaux, E.(2006). Étalonnages à l'aide d'enquêtes de conjoncture: de nouvaux résultats. *Économie et Prévision, INSEE. 172.*

Duflo, M. (1996). Algorithmes stochastiques. *Math. Appl., Springer Verlag, Berlin. 23.*

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation. *Econometrica. 50*, 987-1008.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association 104*, 1430–1439.

Fokianos, K. and Tjøstheim, D. (2011). Log-linear poisson autoregression. *J. Multivariate Anal. 102*, 563–578.

Franke, J. (2010). Weak dependence of functional INGARCH processes. Report in Wirtschaftsmathematik 126, University of Kaiserslautern.

Francq, C. and Zakoian, J.-M. (2010). GARCH Models: Structure, Statistical Inference and Financial Applications. *J. Multivariate Anal.*, Wiley-Blackwell.

Gerchinovitz, S. (2011). Sparsity regret bounds for individual sequences in online linear regression. *Proceedings of COLT'11.*

Gordin, M. I. (1969). The central limit theorem for stationary processes. *Dokl. Akad. Nauk SSSR. 188*, 739Ð741.

Goldstein, S. (1979). Maximal coupling. *Z. Wahrsch. verw. Gebiete. 46*, 193Ð204.

Hamilton, J. (1994). Time Series Analysis. *Princeton University Press.*

Hang, H. and Steinwart, I. (2012). Fast learning from α-mixing observations. *Technical report, Fakultät für Mathematik und Physik, Universität Stuttgart.*

Higgs, M. and Shawe-Taylor, J. (2010). A PAC-Bayes bound for taylored density estimation. *Proceedings of ALT'10, M. Hutter, F. Stephan, V. Vovk and T. Zeugmann Eds. LNAI, Springer.*

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics. 12 no. 1*, 55-67.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and its Application. 7*, 349-382.

Juditsky, A. B. and Nazin, A. V. and Tsybakov, A. B. and Vayatis, N.(2005). Recursive Aggregation of Estimators by the Mirror Descent Algorithm with Averaging. *Methods of Signal Processing. 41*, 368-384.

Juditsky, A. B. and Rigollet, P. and Tsybakov, A. B.(2012). Learning my Mirror Averaging. *Annals of Statistics. 36*, 2183-2206.

Kachour, M. andTruquet, L. (2011). A p-order signed integer-valued autoregressive (SINAR(p)) model. *J. Time Series Anal. 32*, 223–236.

Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis.* Hoboken, NJ: Wiley.

Koenker, R. and Bassett, G. Jr. (1978). Regression quantiles. *Econometrica. 46*, 33-50.

Koenker, R. (2005). Quantile Regression. *Cambridge University Press, Cambridge*.

Kolmogorov, A. N. and Rozanov, Y. A. (1978). On the strong mixing conditions for stationary Gaussian sequences. *Th. Probab. Appl. 5*, 204-207.

Kullback, S. (1959). Information theory and statistics. *Wiley, New York*.

Latour, A. and Truquet, L. (2008). An integer-valued bilinear type model.. *Available at: http://hal.archives-ouvertes.fr/hal-00373409/fr/.*

Li, X. (2010). Agrégation de prédicteurs appliquée à la conjoncture, *Rapport de stage de M2 - Université Paris 6 - INSEE sous la direction de Matthieu Cornec*.

Lecué, G. (2011). Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis. *HDR Thesis, Université Paris-Est Marne-la-Vallée*.

Littlestone, N. and Warmuth, M.K. (1994). The weighted majority algorithm. *Information and Computation. 108*, 212-261.

Nemirovski, A. (2000). Topics in Nonparametric Statistics. *Lectures on Probability Theory and Statistics - Ecole d'ét'e de probabilités de Saint-Flour XXVIII, Springer 2000*, 85-277.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics 34*, 1808-1829.

Massart, P. (2007). Concentration Inequalities and Model Selection - Ecole d'Été de Probabilités de Saint-Flour XXXIII - 2003. *Lecture Notes in Mathematics - J. Picard Editor, Springer* .

Minodier, C. (2010). Avantages comparés des séries premières valeurs publiées et des séries des valeurs révisées. *Documents de Travail de la DESE*.

Modha, D. S. and Masry, E. (1998). Memory-Universal Prediction of Stationary Random Processes. *IEEE transactions on information theory 44*, 117-133.

McAllester, D. A. (1999). PAC-Bayesian Model Averaging. *Procs. of of the 12th Annual Conf. On Computational Learning Theory, Santa Cruz, California (Electronic), ACM, New-York, 1999.*

Mc Leish, D. L. (1975). A generalization of martingales and mixing sequences. *Adv. in Appl. Probab. 7-2*,247-258.

Mc Leish, D. L. (1975a). A maximal inequality and dependent strong laws. *Ann. Probab. 3*,829-839.

Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine Learning 39*, 5-34.

Meyn, S. P. and Tweedie, R. L. (1993). Markov chains and stochastic stability. *Communications and Control Engineering Series, Springer-Verlag London Ltd* .

Modha, D. S. and Masry, E. (1998). Memory-Universal Prediction of Stationary Random Processes. *IEEE transactions on information theory 44*, 117-133.

Neumann, M. H. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli 17*, 1268–1284.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.*

Rakhlin, A. and Sridharan, K. and Tewari, A. (2012). On Empirical Processes with Dependent Data. *http* : *//www* − *stat.wharton.upenn.edu/ rakhlin/papers/emp*$_p$*roc*$_d$*ep.pdf*.

Rio, E. (2000). Théorie asymptotique pour des processus aléatoires faiblement dépendants. *Number* **31** *in Mathématiques et Applications. Springer-Verlag.*

Rio, E. (2000). Ingalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences de Paris, Série I., 330* 905-908.

Robert, C. P. (1996). Méthods de Monte Carlo par chaines de Markov. *Economica (Paris).*

Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal. 4*, 185–207.

Rosenblatt, M. (1956). A central limit Theorem and a strong mixing condition. *Proc. Nat. Ac. Sc. U.S.A. 42* 43-47.

Rosenblatt, M. (1985). Stationary processes and random fields. *Boston: Birkhäuser*.

Salmon, J. and Le Pennec, E. (2009). An aggregator point of view on NL-Mean. *Proceedings of the SPIE Optics and Photonics 2009 Conference on Mathematical Methods: Wavelet XIII, volume 7446, SPIE*, 74461E.

Sanchez-Perez, A. (2013). Time series prediction via aggregation : an oracle bound including numerical cost. *Preprint arXiv:1311.4500*.

Samson, P.-M. (2000). Concentration of measure inequalities for markov chains and Φ-mixing processes. *The Annals of Probability*, *28* 426-461.

Seldin, Y. and Laviolette, F. and Cesa-Bianchi, N. and Shawe-Taylor, J. and Peters, J. and Auer, P. (2012). PAC-Bayesian Inequalities for Martingales. *IEEE Transactions on Information Theory* , *58, no. 12* , 7086-7093.

Shawe-Taylor, J. and Williamson, R. (1997). A PAC Analysis of a Bayes Estimator. *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT'97, ACM,* 2-9.

Steinwart, I. and Anghel, M. (2009). An SVM approach for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *Annals of Statistics*, *37* 841-875.

Steinwart, I. and Christmann, A.(2009). Fast learning from non-i.i.d. observations. *Advances in Neural Information Processing Systems 22* , 1768-1776.

Steinwart, I. and Hush, D. and Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis. 100*, 175-194.

Stoltz, G. (2009). Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la SFDS. 151, no. 2*, 66-106.

Taleb, N. N. (2009). Black Swans and the Domains of Statistics. *The American Statistician. 61*, 198-200.

Tay, A. S. and Wallis, K. F. (2009). Density forecasting: a survey. *Journal of Forecasting. 19*, 258-254.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological. 58*, 267–288.

Tsybakov, A. (2003). Optimal rates of Aggregation. *Learning Theory and Kernel Machines(Schölkopf, B. and Warmuth, M. K.) ,Springer LNCS*, *2003* 303-313.

Vapnik, V. (1999). The nature of statistical learning theory. *Springer*.

Vovk, V.(2001). Competitive on-line statistics. *International Statistical Review. 69*, 218-248.

Vovk, V. G. (1990). Aggregating strategies. *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT).* , 372-383.

Wallis, K. F. (2003). Chi-squared Tests of Interval and Density Forecasts, and the Bank of England's Fan Charts. *International Journal of Forecasting. 19*, 163-175.

Wintenberger, O. (2010). Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability. 15*, 489-503.

Wolkonski, V. A., Rozanov, Y. A. (1959, 1961). Some limit theorems for random functions. Part I: *Theory Probab. Appl. 4*, 178-197; Part II: *Theory Probab. Appl. 6*, 186-198.

Wu, W. B.(2005) . Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences USA. 102*, 14150–14154.

Wu,W. B. and Shao, Z.(2007) . Inference of trends in time series. *J. R. Statist. Soc. B 69*, 391–410.

Wu, W. B. (2007) . Strong invariance principles for dependent random variables. *Ann. Probab. 35*, 2294- 2320.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and Its Interface 4*, 207–226.

Wu, W. B. and Shao, X. (2004). Limit Theorems for Iterated Random Functions. *J. Appl. Probab 41*, 425–436.

Wu, W. B. and Zhou, Z(2011) . Gaussian approximations for non-stationary multiple time series. *Statistica Sinica 21*, 1397- 1413.

Xu, Y.-L. and Chen, D.-R.(2008) . Learning rate of regularized regression for exponentially strongly mixing sequence. *Journal of Statistical Planning and Inference, 138*, 2180-2189.

Zheng, H., Basawa, I. V. and Datta, S. (2006). Inference for the pth–order random coefficient integer–valued process. *Journal of Time Series Analysis 27*, 411–440.

Zheng, H., Basawa, I. V. and Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference. 137*, 212–229.

Zou, B. and Li, L. and Xu, Z.(2009) . The generalization performance of ERM algorithm with strongly mixing observations. *Machine Learning 75*, 275-295.