Growth dynamics of large networks using hidden Markov chains

Quentin Duchemin

Université Gustave Eiffel

Thèse de doctorat
de l'Université Gustave Eiffel

École Doctorale n°532, Mathématiques et STIC (MSTIC)

Spécialité : Mathématiques Appliquées

Thèse préparée au sein du
Laboratoire d'analyse et de mathématiques appliquées (LAMA)

Soutenue publiquement le 20 juin 2022 par

Quentin DUCHEMIN

Dynamique de croissance de grands réseaux à l'aide de chaînes de Markov cachées

---

Devant le jury composé de :

M. Sébastien Bubeck	Microsoft Research	Examinateur
M. Yohann De Castro	Ecole Centrale de Lyon	Co-directeur
Mme Olga Klopp	ESSEC Business School	Rapportrice
Mme Claire Lacour	Université Gustave Eiffel	Directrice
Mme Catherine MATIAS	Sorbonne Université	Examinatrice
M. Nicolas Verzelen	Université de Montpellier	Examinateur
M. Pierre-André Zitt	Université Gustave Eiffel	Examinateur

Après avis des rapporteurs :

M. Charles Bordenave	Institut de Mathématiques de Marseille
Mme Olga Klopp	ESSEC Business School

Remerciements

Mes premiers mots vont naturellement à mes deux encadrants de thèse. Je souhaite remercier Yohann pour son soutien, son temps et sa gentillesse. Mes nombreuses visites à Lyon ont été des étapes clés dans ma thèse au cours desquelles son enthousiasme et sa vision mathématique hors du commun m'ont permis d'avancer sur nos différents projets. Un immense merci à Claire pour sa rigueur et son expertise qui m'ont beaucoup apporté. Sans sa bienveillance et son soutien, cette thèse n'aurait pas été la même.

Je remercie Olga Klopp and Charles Bordenave qui m'ont fait l'honneur de rapporter mon manuscrit de thèse. Je suis également reconnaissant envers les autres membres de jury qui ont accepté de participer à ma soutenance. Mon aventure doctorale a été principalement rythmée par la lecture des travaux de mes rapporteurs et des autres membres du jury. Je mesure donc le privilège qui m'est fait de pouvoir exposer mon travail devant des scientifiques d'un tel calibre.

Cette thèse n'aurait pas été possible sans avoir un goût prononcé pour les mathématiques. Cette passion, je la dois avant tout aux nombreux professeurs d'exception qui ont jalonnés mon parcours. Merci à Monsieur Secouard et Monsieur Lesellier du lycée Fresnel et merci à Monsieur Schweitzer du lycée Malherbe pour qui j'ai un profond respect et qui me fait l'honneur aujourd'hui de son amitié. Je remercie enfin les nombreux enseignants chercheurs de l'Ecole des Ponts pour leur investissement remarquable dans les enseignements de la formation d'ingénieur en IMI. Je pense tout particulièrement à trois enseignants. Le premier est Jean-François Delmas que je remercie profondément pour le temps qu'il m'a accordé ces dernières années en suivant avec intérêt mon parcours et pour m'avoir conseillé et aiguillé dans mes choix. Le second est Guillaume Obonzinski que je remercie de m'avoir accueilli en stage en 2018 et de m'avoir fait rencontrer Yohann. Enfin je souhaiterais remercier Julien Reygner qui m'a donné l'opportunité d'enseigner le cours *Statistiques et Analyse de Données* à l'Ecole des Ponts pour les étudiants de deuxième année. Les qualités humaines et scientifiques font de l'ensemble des chercheurs du CERMICS et du LIGM des personnes d'exception. Ils resteront pour moi de véritables sources d'inspiration.

Cette thèse a aussi été rythmée par une collaboration avec le *Center for Data-Science* de l'université de New-York, initiée à l'occasion de mon stage de fin d'études. J'ai eu la chance de continuer à travailler sur ce projet durant la quasi-totalité de ma thèse. A ce titre, je remercie Carlos Fernandez-Granda et Jakob Assländer pour tout le temps qu'il m'ont consacré et pour leur confiance. Nous avons réalisé ensemble ces dernières années un nombre significatif de visio-conférences au cours desquelles j'ai beaucoup appris, aussi bien sur les aspects techniques liés à l'apprentissage profond que sur les phénomènes physiques et biologiques à l'oeuvre dans les méthodes d'imagerie médicale par résonance magnétique.

Merci également à l'ensemble du personnel de l'UGE, de Centrale Lyon et de l'Ecole Doctorale MSTIC pour leur accompagnement dans toutes les démarches administratives et leur gentillesse. Merci notamment à Isabelle Dominique, Audrey Patout, Gaëlle Lissorgues, Sylvie Cach, Mariam Sidibé et Ketty Cimonard.
Je remercie Hélène, Jean, Ahmed, Benjamin, Josué, Dylan, Michel, Arafat, Kacem, Toan, Martin, Maxime, Gayane et Eddy pour nos discussions et nos débats toujours animés et passionnants autour de nos sujets respectifs mais pas que ! Enfin je voudrais remercier Elias pour son dynamisme, son soutien, sa bonne humeur et tous les très bons moments partagés pendant cette thèse.
Je n'oublie pas bien sûr tous mes très bons amis rencontrés avant cette aventure doctorale et qui me font encore aujourd'hui l'honneur de leur attention malgré l'écart géographique. Je pense notamment à Pierre, Marius, Antoine, Louis, Mehdi, William et Yonatan. Il m'est impossible de ne pas mentionner Anaïs dont le soutien dépasse largement le cadre temporel de la thèse et à qui je dois beaucoup.
Enfin, je remercie chaleureusement tous les membres de ma famille. Ma gratitude va à mes parents et ma soeur pour leur soutien inconditionnel. Les valeurs qui caractérisent leur comportement quotidien constituent pour moi une forme d'idéal à atteindre.

Je remercie le DIM Maths Innov de m'avoir attribué une bourse pour réaliser ces trois années de doctorat. Je remercie l'Ecole Centrale de Lyon d'avoir pu financer plusieurs visites à l'institut Camille Jordan pour travailler avec mon co-directeur de thèse Yohann De Castro. Enfin je remercie l'école doctorale MSTIC ainsi que le laboratoire LAMA de l'UGE d'avoir pu financer des participations à des conférences internationales et à des écoles de printemps qui ont été des expériences marquantes de ma thèse.

Abstract

The first part of this thesis aims at introducing new models of random graphs that account for the temporal evolution of networks. More precisely, we focus on growth models where at each instant a new node is added to the existing graph. We attribute to this new entrant properties that characterize its connectivity to the rest of the network and these properties depend only on the previously introduced node. Our random graph models are thus governed by a latent Markovian dynamic characterizing the sequence of nodes in the graph. We are particularly interested in the Stochastic Block Model and in Random Geometric Graphs for which we propose algorithms to estimate the unknown parameters or functions defining the model. We then show how these estimates allow us to solve link prediction or collaborative filtering problems in networks.
The theoretical analysis of the above-mentioned algorithms requires advanced probabilistic tools. In particular, one of our proof is relying on a concentration inequality for U-statistics in a dependent framework. Few papers have addressed this thorny question and existing works consider sets of assumptions that do not meet our needs. Therefore, the second part of this manuscript will be devoted to the proof of a concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. In Chapter 5, we exploit this concentration result for U-statistics to make new contributions to three very active areas of Statistics and Machine Learning.
Still motivated by link prediction problems in graphs, we study post-selection inference procedures in the framework of logistic regression with L^1 penalty. We prove a central limit theorem under the distribution conditional on the selection event and derive asymptotically valid testing procedures and confidence intervals.

Keywords: Random Graphs, Markov chains, Non-parametric Estimation, Concentration of measure, Integral Operators, Post-selection Inference, Online Learning.

Résumé

La première partie de cette thèse vise à introduire de nouveaux modèles de graphes aléatoires rendant compte de l'évolution temporelle des réseaux. Plus précisément, nous nous concentrons sur des modèles de croissance où à chaque instant un nouveau noeud s'ajoute au graphe existant. Nous attribuons à ce nouvel entrant des propriétés qui caractérisent son pouvoir de connectivité au reste du réseau et celles-ci dépendent uniquement du noeud précédemment introduit. Nos modèles de graphes aléatoires sont donc régis par une dynamique markovienne latente caractérisant la séquence de noeuds du graphe. Nous nous intéresserons particulièrement au Stochastic Block Model et aux Graphes Aléatoires Géométriques pour lesquels nous proposons des algorithmes permettant d'estimer les paramètres du modèle. Nous montrons ensuite comment ce travail d'estimation nous permet de résoudre des problèmes de prédiction de lien ou de filtrage collaboratif dans les graphes.
L'étude théorique des algorithmes précédemment décrits mobilisent des résultats probabilistes poussés. Nous avons notamment dû recourir à une inégalité de concentration pour les U-statistiques dans un cadre dépendant. Peu nombreux sont les travaux ayant abordé cette épineuse question et l'existant considère des jeux d'hypothèses ne répondant pas à nos besoins. Aussi, la deuxième partie de ce manuscrit sera consacrée à la preuve d'une inégalité de concentration pour les U-statistiques d'ordre deux pour des chaînes de Markov uniformément ergodique. Dans le Chapitre 5, nous exploitons notre résultat de concentration pour les U-statistiques pour apporter de nouvelles contributions à trois domaines très actifs des Statistiques et du Machine Learning.
Toujours motivés par des problèmes de prédictions liens dans les graphes, nous nous intéressons dans un dernier chapitre aux procédures d'inférence post-sélection dans le cadre de la régression logistique avec pénalité L^1. Nous prouvons un théorème central limite sous la distribution conditionnelle à l'événement de sélection et nous en déduisons des procédures de test et des intervalles de confiance asymptotiquement valides.

Mots-clés : Graphes aléatoires, Chaines de Markov, Estimation non-paramétrique, Concentration de la mesure, Opérateurs intégraux, Inférence post-sélection, Apprentissage séquentiel.

List of Publications

- Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. Markov random geometric graph, MRGG: A growth model for temporal dynamic networks. *Electron. J. Stat.*, 16 (1):671–699, 2022. doi: 10.1214/21-ejs1969. URL https://doi.org/10.1214/21-ejs1969
- Duchemin et al. [2022b] Quentin Duchemin, Yohann De Castro, and Claire Lacour. Concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. *Bernoulli*, 2022b. URL https://hal.archives-ouvertes.fr/hal-03014763
- Duchemin et al. [2022a] Quentin Duchemin, Yohann De Castro, and Claire Lacour. Three rates of convergence or separation via U-statistics in a dependent framework. *JMLR*, 2022a. URL https://hal.archives-ouvertes.fr/hal-03603516
- Zhang et al. [2022] Xiaoxia Zhang, Quentin Duchemin, Kangning Liu, Sebastian Flassbeck, Cem Gultekin, Carlos Fernandez-Granda, and Jakob Assländer. Cramér-Rao bound-informed training of neural networks for quantitative MRI. *Magnetic Resonance in Medicine*, 2022. doi: https://doi.org/10.1002/mrm.29206. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.29206
- Duchemin et al. [2020] Quentin Duchemin, Kangning Liu, Carlos Fernandez-Granda, and Jakob Assländer. Optimized dimensionality reduction for parameter estimation in MR fingerprinting via deep learning. *ISMRM*, 3750(1):189–195, 2020

List of Preprints

- Duchemin [2022] Quentin Duchemin. Reliable Time Prediction in the Markov Stochastic Block Model. preprint, March 2022. URL https://hal.archives-ouvertes.fr/hal-02536727
- Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. The Random Geometric Graph: Recent developments and perspectives. 2022. URL https://hal.archives-ouvertes.fr/hal-03622277
- Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. A new procedure for Selective Inference with the Generalized Linear Lasso. 2022. URL https://hal.archives-ouvertes.fr/hal-03622196

Code

- Reliable time predictions in MSBMs (Chapter 2): https://github.com/quentin-duchemin/inference-markovian-SBM.
- Non-parametric estimation toolbox and link prediction for MRGGs (Chapter 3): https://github.com/quentin-duchemin/Markovian-random-geometric-graph.
- Estimation of biomarkers in Magnetic Resonance Fingerprinting: https://quentin-duchemin.github.io/MRF-CRBLoss/index.html.
- Post-selection inference in the logistic model (Chapter 6): https://github.com/quentin-duchemin/LogPSI.

Contents

List of Figures

List of Tables

Notations

$\mathbb{N}^*$	Set of positive natural integers $\mathbb{N} \setminus \{0\}$.
$[n]$	Set of integers $\{1, \dots, n\}$.
$\lvert A \rvert$	Cardinality of any finite set A.
A^c	Complement of any set A.
$\delta_{i,j}$	For any integers i, j, the Kronecker symbol $\delta_{i,j}$ is equal to 1 if $i = j$ and is equal to 0 otherwise.
$\mathbf{1}_n$	The all-ones vector of size n.
$\mathfrak{S}_N$	The set of permutations of $[N]$.
$x \vee y$	Maximum between reals x and y.
$x \wedge y$	Minimum between reals x and y.
$\lceil x \rceil$	The smallest integer larger than x.
$\lfloor x \rfloor$	The largest integer less than x.
$a_n = \mathcal{O}(b_n)$ or $b_n = \Omega(a_n)$	Given two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ of real numbers such that for some $N \in \mathbb{N}$, $b_n \neq 0$ for all $n \geq N$, we write $a_n = \mathcal{O}(b_n)$ or $b_n = \Omega(a_n)$ if the sequence $(a_n/b_n)_{n\geq N}$ is bounded and we write $a_n = o(b_n)$ or $b_n = \omega(a_n)$ if $a_n/b_n \underset{n\to+\infty}{\to} 0$.
$\mathbb{S}^{d-1}$	For any integer $d \geq 2$, the Euclidean Sphere of dimension d is given by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$.
ℓ_2	The Hilbert space of all square summable sequences.
$L^p(\pi)$	For any measure π on some measurable space (E, Σ) and for any $p \in [1, \infty)$, we denote by $L^p(\pi)$ the space of measurable functions $h : E \to \mathbb{R}$ for which the p-th power of the absolute value is π-integrable, where functions which agree π-almost everywhere are identified. $L^\infty(\pi)$ is the space of measurable functions $h : E \to \mathbb{R}$ bounded π-almost everywhere.
$\mathcal{B}(\mathbb{R})$	The Borel algebra on $\mathbb{R}$.
$\|x\|_p$	For any $p \geq 1$, the ℓ_p norm of any vector $x \in \mathbb{R}^n$ is defined by $\|x\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ while $\|x\|_\infty = \max_{i\in[n]} \lvert x_i \rvert$.
$\langle x, y \rangle$	The Euclidean inner product on $\mathbb{R}^d$ is denoted by $\langle \cdot, \cdot \rangle : (x, y) \mapsto \langle x, y \rangle = \sum_{i=1}^d x_i y_i$.
$M^\top$, $M_{i,:}$ and $M_{:,i}$	Matrices are denoted in standard font capital letters (M). The transpose of a matrix is $M^\top$. The i-th row and column of a matrix M are denoted with $M_{i,:}$ and $M_{:,i}$, respectively. $M_{i,j}$ corresponds to the entry of M at row i and column j.
$\|M\|$	The operator norm of a matrix $M \in \mathbb{R}^{n\times d}$ is defined by $\|M\| = \sup_{x\in\mathbb{S}^{d-1}} \|Mx\|_2$.
$\|M\|_F$	The Frobenius norm of a matrix $M \in \mathbb{R}^{n\times d}$ is defined by $\|M\|_F = \left(\sum_{i\in[n]} \sum_{j\in[d]} M_{i,j}^2\right)^{1/2}$.
$\mathrm{Tr}(M)$	The trace of a matrix $M \in \mathbb{R}^{n\times n}$, *i.e.* $\mathrm{Tr}(M) = \sum_{i=1}^n M_{i,i}$.
$\odot$	Denotes the Hadamard product namely for any $A, B \in \mathbb{R}^{d\times p}$, $A \odot B := (A_{i,j} B_{i,j})_{i\in[d], j\in[p]}$.
$\lambda(M)$	For any matrix $M \in \mathbb{R}^{n\times n}$, $\lambda(M)$ denotes the set of eigenvalues of M.
$\delta_2(x, y)$	Given two sequences x, y of reals–completing finite sequences by zeros–such that $\sum_i x_i^2 + y_i^2 < \infty$, we define the ℓ_2 rearrangement distance $\delta_2(x, y)$ as $$\delta_2^2(x, y) := \inf_{\sigma\in\mathfrak{S}} \sum_i (x_i - y_{\sigma(i)})^2 ,$$ where $\mathfrak{S}$ is the set of permutations with finite support. This distance is useful to compare two spectra.

Id_d	The diagonal identity matrix of size $d \times d$.
$\mathcal{B}(\alpha, \beta)$	For any $\alpha, \beta > 0$, $\mathcal{B}(\alpha, \beta)$ will denote the beta distribution with α and β.
$\mathcal{N}(\mu, \Sigma)$	For any $\mu \in \mathbb{R}^d$ and any pseudo-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and covariance matrix Σ.
$\mathrm{Ber}(p)$	For any $p \in [0, 1]$, $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p.
$\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}$	Considering two measures π, χ on some measurable space (E, Σ) such that χ is absolutely continuous with respect to π. For $p \in [1, \infty]$, the p-th moment of the density of χ with respect to π - denoted $\frac{d\chi}{d\pi}$ - is defined by $\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} := \begin{cases} \left[\int \left|\frac{d\chi}{d\pi}\right|^p d\pi\right]^{1/p} & \text{if } p < \infty, \\ \operatorname{ess\,sup} \left|\frac{d\chi}{d\pi}\right| & \text{if } p = \infty. \end{cases}$
$\overset{(d)}{=}$	Denotes the equality in distribution sense.
$\overset{(d)}{\to}$	Denotes the weak convergence.
$\|\omega\|_{\mathrm{TV}}$	For any probability measure ω on some measurable space (E, Σ), $\|\omega\|_{\mathrm{TV}} := \sup_{A \in \Sigma} \|\omega(A)\|$ is the total variation norm of ω.
$X_n = \mathcal{O}_{\mathbb{P}}(a_n)$	Given a sequence of real valued random variables $(X_n)_{n \in \mathbb{N}}$ and a sequence of positive reals $(a_n)_{n \in \mathbb{N}}$, the notation $X_n = \mathcal{O}_{\mathbb{P}}(a_n)$ means that $(X_n / a_n)_{n \in \mathbb{N}}$ converges to zero in probability as $n \to \infty$.
$\mathcal{N}(\mathcal{H}, \eta)$	For any $\eta > 0$, $\mathcal{N}(\mathcal{H}, \eta)$ is the L^∞ *η-covering number* of the set $\mathcal{H}$.
$\sigma(X_i,\, i \in I)$	Given $I \subset \mathbb{N}$ and random variables $(X_i)_{i \in I}$, we denote by $\sigma(X_i,\, i \in I)$ the σ-algebra generated by the random variables X_i, $i \in I$.
"w.h.p."	A sequence of events $(A_n)_{n \in \mathbb{N}}$ is said to hold with high probability (we will use the abbreviation "w.h.p."), if $\mathbb{P}(A_n)$ converges to 1 when $n \to \infty$.

Abbreviations

SDP	Semi Definite Programming
SNR	Signal to Noise Ratio
SBM	Stochastic Block Model
MSBM	Markov Stochastic Block Model
HAC	Hierarchical Agglomerative Clustering
HEiC	Harmonic Eigen Cluster
SCCHEi	Size Constrained Clustering for Harmonic Eigenvalues
RGG	Random Geometric Graph
MRGG	Markov Random Geometric Graph
PSI	Post-Selection Inference
SLR	Sparse Logistic Regression
SEI-SLR	Selection Event Identification for Sparse Logistic Regression
GLM	Generalized Linear Model
GLL	Generalized Linear Lasso
MCMC	Monte Carlo Markov Chain
MLE	Maximum Likelihood Estimator
CLT	Central Limit Theorem
SA	Simulated Annealing
IRLS	Iteratively Reweighted Least Squares
KKT	Karush Kuhn Tucker

Résumé de la thèse

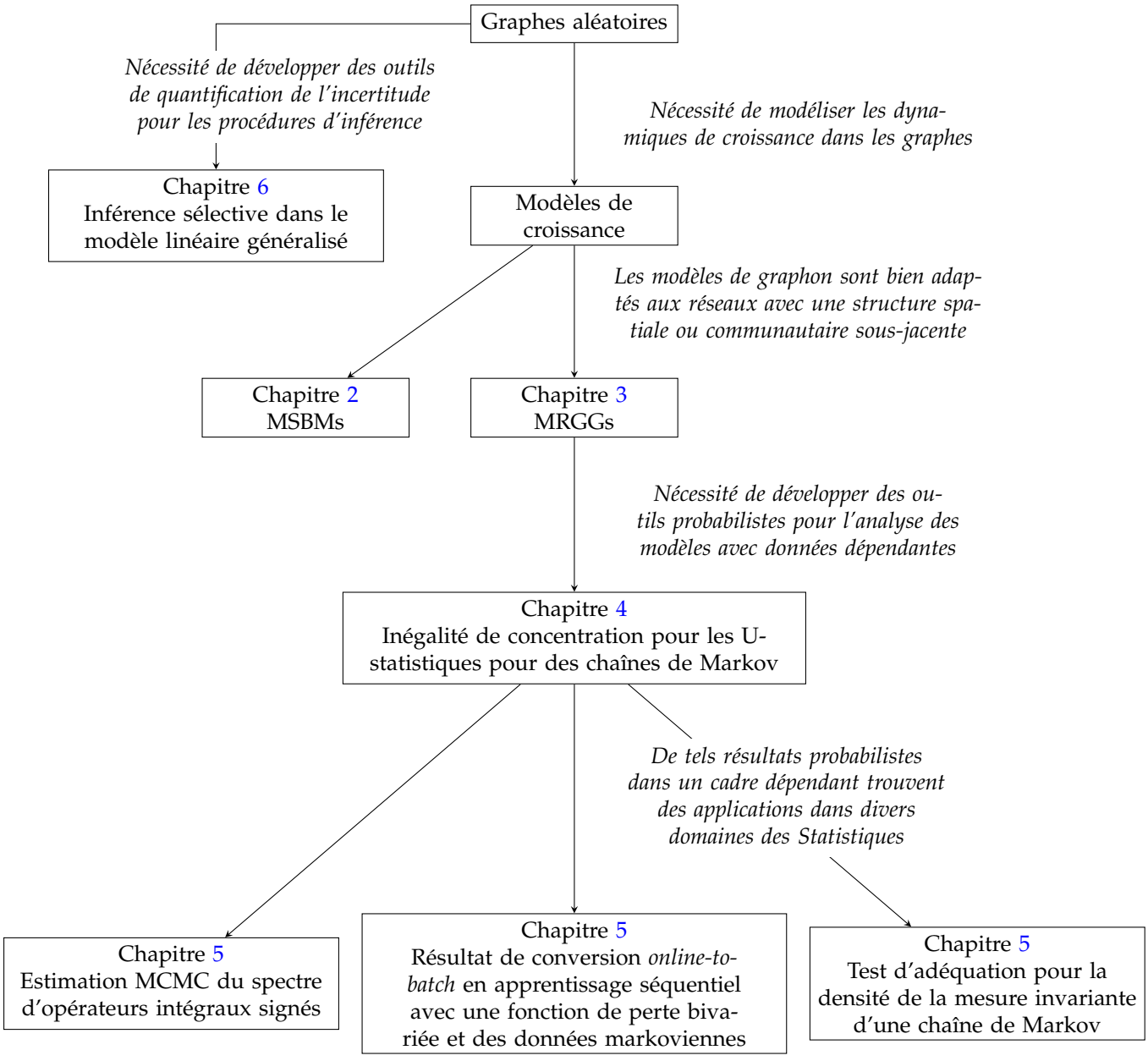


FIGURE 1 : Structure de la thèse.

1. Modèles de croissance dans les graphes aléatoires

1.1 Modélisation de réseaux par des graphes aléatoires

Graphes aléatoires. Les graphes sont aujourd'hui largement utilisés pour modéliser des systèmes complexes dans les applications réelles. Comme il s'agit d'objets de grande dimension, il est nécessaire de supposer une certaine structure sur les données pour pouvoir extraire efficacement des informations sur le système étudié. À cette fin, un grand nombre de modèles de graphes aléatoires ont déjà été introduits. Le plus simple est le modèle d'Erdös-Renyi $G(n, p)$ dans lequel chaque arête entre des paires de n nœuds est présente dans le graphe avec probabilité $p \in (0, 1)$. On peut également mentionner le *scale-free network model* de [Barabási, 2009] ou les *small-world networks* de [Watts and Strogatz, 1998]. Nous renvoyons à [Channarond, 2015] pour une introduction aux modèles de graphes aléatoires les plus célèbres. En pratique, il existe souvent des variables pertinentes expliquant l'hétérogénéité des observations. La plupart du temps, ces variables explicatives sont inconnues et portent une information précieuse sur le système étudié. Dans un tel contexte, les modèles à espace latent sont des outils bien adaptés pour représenter les données (voir Smith et al. [2019]). Parmi les modèles à espace latent les plus étudiés, on trouve ceux construits à partir de *communautés cachées* où chaque nœud est supposé appartenir à un (ou plusieurs) groupe(s) tandis que les probabilités de connexion entre deux nœuds du graphe dépendent de leur appartenance respective. Le célèbre *Stochastic Block Model* (SBM) a fait l'objet d'une attention particulière ces dernières années. Nous renvoyons à Abbe [2017] pour une excellente introduction à ce modèle et aux questions statistiques et algorithmiques en jeu. Dans les modèles d'espace latent mentionnés précédemment, la géométrie intrinsèque du problème n'est pas prise en compte. Cependant, la structure spatiale sous-jacente des graphes est essentielle puisque la géométrie affecte considérablement la topologie de ces derniers (voir Barthélemy [2011] et Smith et al. [2019]). Afin de répondre à ce besoin de modélisation, des graphes aléatoires avec des structures latentes continues plus complexes que celles des SBMs ont été étudiés, comme les *Random Geometric Graphs* (RGGs).

Les modèles de graphon : les cas particuliers des SBMs et des RGGs. Les modèles du SBM et du RGG sont des cas particuliers de modèle de graphon [voir Lovász, 2012]. Dans un modèle de graphon, nous considérons un espace latent $\mathcal{X}$ et une fonction symétrique $W : \mathcal{X} \times \mathcal{X} \to [0, 1]$. Pour construire un graphe simple et non orienté de taille n à partir du modèle graphon associé à W, il faut d'abord échantillonner les positions latentes $(X_i)_{i\in[n]}$ dans $\mathcal{X}^n$. Deux sommets $i, j \in [n]$ avec $i \neq j$ sont alors connectés avec probabilité $W(X_i, X_j)$.
Les modèles du SBM et du RGG diffèrent essentiellement par leur structure topologique. C'est la raison pour laquelle ils sont utilisés en pratique pour modéliser des phénomènes physiques de natures différentes. En prenant l'exemple des réseaux sociaux, nous pouvons considérer en première approximation [comme expliqué dans Péché and Perchet, 2020] qu'une connexion entre deux utilisateurs peut se produire pour deux raisons principales : (i) soit ces derniers sont des amis d'enfance (ce qui signifie que les utilisateurs sont endogènement similaires avec des représentations géométriques latentes proches) ou (ii) ils partagent les mêmes opinions politiques (ce qui signifie que les utilisateurs sont exogènement similaires et appartiennent à une même communauté). Ces deux modèles sont donc intéressants pour les applications modernes et ont suscité un intérêt grandissant au cours des dernières décennies. Outre cette dimension applicative, les modèles du SBM et du RGG sont par eux-mêmes des objets d'étude particulièrement riches sur le plan mathématique.

Prédictions temporelles. Les réseaux réels évoluent au cours du temps et nous pouvons citer l'exemple de la propagation de maladies [cf. Barthélemy, 2011, Section 5.6.3]. Afin de tendre vers une modélisation plus fidèle de toute la complexité des données réelles, de nombreux chercheurs ont développé des nouveaux modèles de graphes aléatoires rendant compte de la dimension évolutive des réseaux. La plupart d'entre eux étendent les modèles à espace latent et considère un graphe de taille fixe où les arêtes et/ou les représentations latentes peuvent changer au cours du temps [cf. Lei and Rinaldo, 2015, Matias and Miele, 2015, 2017, Xie and Rogers, 2016, Xie et al., 2015, Xu and Hero, 2014]. Dans cette thèse, nous adoptons un tout autre point de vue. Motivés par les problèmes de prédiction de liens dans les graphes, nous nous concentrons sur les *modèles de croissance*, à savoir des modèles de graphes aléatoires dans lesquels à chaque pas de temps, un nouveau nœud rejoint le réseau existant et se connecte aux autres sommets du graphe selon une règle probabiliste qui doit être spécifiée. Au cours de la dernière décennie, les modèles de croissance pour les graphes aléatoires avec une structure spatiale latente ont suscité un intérêt accru. On peut citer Jordan and Wade [2015], Papadopoulos et al. [2012] et Zuev et al.

[2015] où des variantes géométriques du modèle d'attachement préférentiel sont introduites avec un nouveau nœud entrant dans le graphe à chaque pas de temps.
Dans ce qui suit, nous proposons des modèles de croissance basés sur le SBM et sur le RGG où les attributs latents des nœuds sont échantillonnés en utilisant une dynamique markovienne. Nous montrons que ces nouveaux modèles sont pertinents pour résoudre des problèmes de prédiction de liens où il s'agit d'estimer la probabilité de connexion entre les nœuds déjà présents dans le graphe et les futurs arrivants.

1.2 Un modèle de croissance pour le SBM (Chapitre 2)

Présentation du Stochastic Block Model. Le SBM est un terrain de jeu parfait pour étudier l'existence de phénomènes de transition de phase et de *gaps* informationnels/computationnels. Dans le SBM, nous cherchons à obtenir des informations sur les communautés latentes cachées à partir de l'observation de la matrice d'adjacence du graphe. Différents critères statistiques peuvent être étudiés dans les SBM, telles que l'*exact recovery* (où l'on vise à estimer correctement l'ensemble de la partition des noeuds de graphe avec grande probabilité) ou le *weak recovery* aussi appelé problème de détection (où l'algorithme doit fournir avec grande probabilité une partition des noeuds positivement corrélée à la partition cachée). Un grand nombre de méthodes sont aujourd'hui connues pour aborder ces différents problèmes. On peut citer par exemple les méthodes probabilistes basées sur l'algorithme EM, la programmation semi-définie, ou encore les méthodes spectrales. Même si le SBM a été largement étudié, ce modèle mobilise encore une large communauté de chercheurs qui s'adonnent à la résolution de questions théoriques encore ouvertes ou tentent d'étendre le modèle pour permettre une description plus riche des réseaux. Dans ce qui suit, nous mettons en lumière certains de ces enjeux clés.

- ✓ *Le régime parcimonieux.* Lorsque le degré moyen des nœuds du graphe est constant (i.e., ne dépend pas de n), les SBMs ne sont pas connexes avec grande probabilité et il n'est pas possible de résoudre le problème d'*exact recovery*. Dans ce cas, l'objectif est de trouver un algorithme qui résout le problème de détection. La plupart des méthodes utilisées pour résoudre le problème *exact recovery* - telles que les méthodes spectrales sur les matrices laplaciennes du graphe - ne permettent pas de résoudre le problème de détection dans le régime parcimonieux. Dans Krzakala et al. [2013], les auteurs ont introduit une nouvelle représentation matricielle du graphe appelée la matrice *non-backtracking* B et ont affirmé qu'une méthode spectrale sur B pourrait résoudre le problème de détection dans le régime parcimonieux. Cette réhabilitation des méthodes spectrales via l'opérateur *non-backtracking* a été rigoureusement prouvée par Bordenave et al. [2018] pour le SBM symétrique avec deux communautés. La conjecture générale pour un nombre arbitraire de communautés symétriques ou asymétriques a été résolue plus tard dans Abbe and Sandon [2015b] en s'appuyant sur une matrice *non-backtracking* d'ordre supérieur et une implémentation de type *message passing*.

- ✓ *L'hétérogénéité des degrés.* Une autre limite principale des méthodes spectrales pour les applications réelles est lorsque les degrés du réseaux sont hétérogènes. En travaillant avec le *degree correlated SBM*, Dall'Amico and Couillet [2019] ont prouvé que la matrice *Bethe-Hessian* peut être utilisée pour résoudre le problème de détection de communautés dans les graphes parcimonieux à degrés inhomogènes.

- ✓ *Partial recovery bounds.* En interpolant entre les critères d'*exact* et de *weak recovery*, l'un des principaux défis des SBMs est de comprendre le lien inhérent entre un rapport signal/bruit (SNR) approprié et l'erreur de classification, à savoir l'erreur d'alignement entre la partition des noeuds du graphe renvoyée par l'algorithme et la partition cachée. Dans Giraud and Verzelen [2019], les auteurs ont proposé un algorithme de programmation semi-définie positive (SDP) pour estimer les communautés et ont prouvé que l'erreur de classification décroit exponentiellement vite vers 0 pour un SNR bien choisi.

La discussion précédente s'est concentrée sur des graphes statiques. Cependant, dans de nombreuses applications, nous avons accès à plusieurs versions d'un même graphe qui évolue au cours du temps. C'est le cas des réseaux représentant la proximité physique d'agents mobiles ou l'évolution biologique et chimique des membres d'un groupe et nous renvoyons à Holme [2015] pour une revue de la littérature à ce sujet. Afin d'extraire des informations temporelles sur le système d'intérêt, plusieurs travaux ont étendu le SBM pour modéliser la structure dynamique des réseaux étudiés. Dans Matias and Miele [2015], une variante du SBM est considérée où l'évolution temporelle est modélisée par une chaîne de

Markov cachée discrète sur les communautés des nœuds et où les probabilités de connexion évoluent également dans le temps. Inspiré par les travaux de Karrer and Newman [2011], Lei and Rinaldo [2015] étudie le *degree correlated SBM* où le degré des nœuds peut varier au sein d'une même communauté. Nous pouvons également mentionner Keriven and Vaiter [2022] ou Dall'Amico et al. [2020].
Les travaux mentionnés ci-dessus considèrent principalement des SBMs où les communautés des nœuds ou les présences/absences d'arêtes peuvent évoluer dans le temps, mais rares sont les articles qui s'intéressent à des modèles de croissance pour les SBMs. Dans cette thèse, nous cherchons à combler ce manque.

Le Markov Stochastic Block Model. Dans le chapitre 2, nous introduisons le MSBM (Markov SBM) : une extension du SBM où les communautés des noeuds sont alouées via une dynamique markovienne. Notre objectif est de fournir des méthodes efficaces et fiables pour résoudre

- les problèmes de prédiction de liens où nous cherchons à calculer la probabilité de connexion entre les nœuds du graphe et un futur entrant,
- ou les tâches de filtrage collaboratif où nous voulons déduire la communauté cachée d'un nœud si nous n'avons qu'une information partielle sur la façon dont ce nœud est connecté au reste du graphe.

Dans le MSBM, nous considérons qu'à chaque pas de temps i, un nouveau nœud entre dans le réseau avec une communauté latente $C_i \in [K]$ (pour un entier positif fixe K) qui est échantillonnée à partir de la distribution de probabilité $P_{C_i,\cdot}$ où $P \in [0,1]^{K\times K}$ est la matrice de transition de la chaîne de Markov récurrente positive $(C_i)_{i\in[n]}$. Une fois la communauté de chaque nœud attribuée, une arête est créée entre les nœuds i et j avec probabilité Q_{C_i,C_j} où $Q \in [0,1]^{K\times K}$ est la matrice de connectivité. Dans ce modèle, nous pouvons utiliser les méthodes standards développées pour le SBM pour estimer les communautés cachées des nœuds à partir de l'observation de la matrice d'adjacence du graphe. A partir des communautés estimées, nous pouvons fournir des estimations des paramètres du modèle P et Q. Afin de résoudre les problèmes de prédiction de liens ou de filtrage collaboratif, une approche naturelle consiste à s'appuyer sur une approche *plug-in* en utilisant les estimateurs précédemment introduits de P, Q et $(C_i)_{i\in[n]}$. Néanmoins, nous montrons dans le chapitre 2 que cette procédure *plug-in* est très peu robuste car elle est très sensible aux erreurs de clustering faites par l'algorithme utilisé. Dans le chapitre 2, nous proposons une méthodologie générale pour résoudre les tâches de prédiction de liens et de filtrage collaboratif dans le MSBM qui s'adapte aux erreurs locales de clustering et qui s'avère beaucoup plus fiable pour les applications pratiques. Nos contributions sont les suivantes.

- Méthodes générales pour une prédiction temporelle fiable dans les MSBMs.
  $i)$ En établissant un lien entre les MSBMs et les modèles de Markov cachés, nous proposons d'estimer ce que l'on appelle les *probabilités d'émission* qui correspondent pour tout $c, \hat{c} \in [K]$ à la probabilité pour un certain nœud appartenant à la communauté $c \in [K]$ d'être assigné à la communauté $\hat{c}$ par l'algorithme de clustering. Ces quantités nous permettent de concevoir des méthodes fiables de prédiction de liens et de filtrage collaboratif qui peuvent tenir compte des erreurs locales dans les estimations des communautés cachées.
  $ii)$ Nous montrons également comment les probabilités d'émission apprises peuvent être utilisées pour effectuer une sélection de modèle, *i.e.* pour estimer le nombre inconnu de communautés latentes K.
  Nous soulignons que les méthodes proposées peuvent être utilisées pour n'importe quel algorithme de clustering. Afin de mener des expériences numériques et de fournir des garanties théoriques, nous utilisons dans le chapitre 2 l'algorithme de programmation semi-définie (SDP) introduit par Giraud and Verzelen [2019].
- Garanties théoriques.
  $iii)$ Nous identifions un rapport signal sur bruit pertinent dans notre cadre d'étude et nous prouvons que la partition des nœuds obtenue par l'algorithme SDP conduit à une erreur de classification qui décroît exponentiellement vite par rapport à ce signal sur bruit.
  $iv)$ Nous donnons des estimateurs des paramètres du MSBM dont nous prouvons la consistance.
- Aspects numériques.
  $v)$ A notre connaissance, nous sommes les premiers à fournir une implémentation de la méthode SDP de Giraud and Verzelen [2019]. Une partie importante de ce travail a consisté à coder une

étape d'approximation technique issue de Charikar et al. [2002].
vi) Nous fournissons des résultats numériques de nos méthodes sur des données simulées et réelles.

Ce travail sur les MSBMs présenté dans le chapitre 2 correspond à l'article suivant.

Duchemin [2022] Quentin Duchemin. Reliable Time Prediction in the Markov Stochastic Block Model. preprint, March 2022. URL https://hal.archives-ouvertes.fr/hal-02536727

1.3 Un modèle de croissance pour le RGG (Chapitre 3)

Présentation des Random Geometric Graphs. Le modèle RGG a été introduit pour la première fois par Gilbert [1961] pour modéliser les communications entre stations radio. Le modèle original de Gilbert était défini comme suit : choisir des points dans $\mathbb{R}^2$ selon un processus ponctuel de Poisson d'intensité 1 et connecter deux points si leur distance est inférieure à un certain paramètre $r > 0$. Le modèle RGG a été étendu à d'autres espaces latents tels que l'hypercube $[0,1]^d$, la sphère euclidienne ou aux groupes de Lie compacts Méliot [2019]. Une importante littérature a été consacrée à l'étude des propriétés des RGGs en faible dimension Penrose [2003], Dall and Christensen [2002], Bollobás [2001]. Les RGGs ont trouvé des applications dans un très grand nombre de domaines comme les réseaux sans fil Haenggi et al. [2009] ou les algorithmes de bavadarge Wang and Lin [2014].

Pour citer Bollobás [2001], *"Un des principaux objectifs de la théorie des graphes aléatoires est de déterminer quand une propriété donnée est susceptible d'apparaître."* Dans cette direction, plusieurs travaux ont tenté d'identifier la structure des réseaux par une procédure de test, voir par exemple Ghoshdastidar et al. [2020]. En ce qui concerne les RGGs, la plupart des résultats ont été établis en petite dimension $d \leq 3$ [cf. Barthélemy, 2011, Penrose, 2003]. Cependant, l'omniprésence des problèmes statistiques impliquant des données de grande dimension a motivé la communauté à étudier les propriétés des RGGs dans le cas où $d \to \infty$. Le problème qui consiste à déterminer si un graphe peut être réalisé comme un graphe géométrique apparaît comme naturel et important, mais nous savons aujourd'hui qu'il est NP-difficile Breu and Kirkpatrick [1998]. Une question reliée et plus abordable est celle qui consiste à savoir si un RGG donné porte encore de l'information spatiale lorsque $d \to \infty$ ou si la géométrie est perdue en haute dimension. Ce problème, connu sous le nom de *détection de géométrie*, vise à tester si un graphe a été échantillonné à partir d'une distribution Erdös-Renyi ou celle d'un RGG. Cette question a suscité beaucoup d'intérêt au cours des dernières années et nous pouvons mentionner en particulier les contributions importantes de Brennan et al. [2020], Bubeck et al. [2016], Liu et al. [2021]. Les preuves présentées dans ces articles font appel à des résultats avancés issus des probabilités, des statistiques, du transport optimal, de la combinatoire ou de la théorie de l'information, plaçant les RGGs à l'intersection d'un large éventail de communautés de recherche. Nous pouvons mentionner en particulier que dans le régime dense, la transition de phase pour le problème de détection de géométrie se produit au régime auquel les matrices de Wishart deviennent indiscernables des GOE (*Gaussian Orthogonal Ensemble*) pour la distance de variation totale. Cette question a été étudiée dans des contextes plus généraux comme dans Bubeck and Ganguly [2015] ou Bourguin et al. [2021]. Au cours de mon doctorat, j'ai écrit un article de synthèse (qui ne sera pas discuté avec plus de détails dans ce manuscrit) sur les nombreuses questions mathématiques captivantes liées aux RGGs et à leurs extensions.

Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. The Random Geometric Graph: Recent developments and perspectives. 2022. URL https://hal.archives-ouvertes.fr/hal-03622277

Estimation non-paramétrique dans les RGGs. Toujours dans le modèle du RGG, De Castro et al. [2019] aborde une toute autre question et considère un problème d'estimation non-paramétrique dans les RGGs. Leur travail contribue à la question plus large de l'estimation dans les modèles de graphon. Dans Tang et al. [2013], les auteurs prouvent que des méthodes spectrales peuvent permettrer d'estimer la matrice formée par le graphon évalué aux positions latentes à une transformation orthogonale près, en supposant que le graphon est un noyau défini positif. En quittant l'univers discret

et en s'attaquant à l'objet continu sous-jacent, des algorithmes ont été conçus pour estimer les graphons eux-mêmes, comme dans Klopp et al. [2017] qui fournissent des vitesses de convergence optimale pour le SBM. Contrairement aux travaux précédents, l'article De Castro et al. [2019] fournit une approche non-paramétrique pour estimer le graphon caractérisant un RGG sur la sphère euclidienne $\mathbb{S}^{d-1}$, sans faire l'hypothèse que le noyau est défini-positif. Dans leur modèle, les auteurs considèrent n points $X_1, X_2, \dots, X_n$ échantillonnés uniformément et indépendamment sur $\mathbb{S}^{d-1}$ et une arête est créée entre les nœuds i et j (où $i, j \in [n]$, $i \neq j$) avec probabilité $\mathbf{p}(\langle X_i, X_j \rangle)$, où la fonction inconnue $\mathbf{p} : [-1, 1] \to [0, 1]$ est appelée la fonction *enveloppe*. Ce RGG est un modèle de graphon avec un noyau symétrique W donné par $W(x, y) = \mathbf{p}(\langle x, y \rangle)$. On peut associer au graphon W un opérateur intégral $\mathbb{T}_W$. L'opérateur $\mathbb{T}_W$ est de Hilbert-Schmidt et il possède un nombre dénombrable de valeurs propres bornées $\lambda(\mathbb{T}_W)$ avec zéro comme seul point d'accumulation. Les fonctions propres $(\phi_k)_{k \geq 0}$ de $\mathbb{T}_W$ ont la propriété remarquable de ne pas dépendre de $\mathbf{p}$ (cf. Dai and Xu [2013] Lemma 1.2.3) : elles sont données par les harmoniques sphériques réelles. On peut montrer que la décomposition spectrale suivante est vérifiée

$$\mathbf{p}(t) = \sum_{l \geq 0} p_l^* \phi_k(t) \,,$$

où $\lambda(\mathbb{T}_W) = \{p_0^*, p_1^*, \dots, p_1^*, \dots, p_l^*, \dots, p_l^*, \dots\}$, *i.e.* chaque valeur propre p_l^* a une multiplicité connue d_l. Cette décomposition montre qu'une estimation par une approche *plug-in* de la fonction enveloppe $\mathbf{p}$ est possible si l'on est capable d'estimer les valeurs propres de l'opérateur $\mathbb{T}_W$. Dans De Castro et al. [2019], les auteurs prouvent que sous une certaine condition de régularité sur la fonction enveloppe, le spectre de la matrice d'adjacence du graphe (correctement normalisé) converge vers $\lambda(\mathbb{T}_W)$ au sens de la métrique δ_2 définie pour toutes suites de réels x et y de carré sommable par

$$\delta_2^2(x, y) := \inf_{\sigma \in \mathfrak{S}} \sum_i (x_i - y_{\sigma(i)})^2 \,,$$

où $\mathfrak{S}$ est l'ensemble des permutations de $\mathbb{N}$.
Cette approche présente deux inconvénients majeurs : (i) Premièrement, l'algorithme proposé pour estimer la fonction enveloppe a une complexité qui croît de façon exponentielle avec le niveau de résolution choisi R. (ii) Enfin, les auteurs travaillent avec la condition restrictive d'un échantillonnage indépendant des positions latentes.

Le Markov Random Geometric Graph. Dans notre article

> Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. Markov random geometric graph, MRGG: A growth model for temporal dynamic networks. *Electron. J. Stat.*, 16 (1):671–699, 2022. doi: 10.1214/21-ejs1969. URL https://doi.org/10.1214/21-ejs1969

nous cherchons à résoudre les deux problèmes précédents en introduisant le Markov RGG (MRGG) : un modèle de croissance pour les RGGs sur $\mathbb{S}^{d-1}$ où les positions latentes sont échantillonnées en utilisant une dynamique markovienne. Plus précisément, la distribution du nouveau point latent X_i est donnée par $P(X_{i-1}, \cdot)$ où P est le noyau de transition de la chaine de Markov $(X_i)_{i \geq 1}$ qui doit être estimé. Nous considérons un schéma d'échantillonnage markovien isotrope, ce qui signifie qu'à partir d'une position latente X_i, le point latent suivant est défini par

$$X_i = r_i \times X_{i-1} + \sqrt{1 - r_i^2} \times Y_i,$$

où

- $Y_i \in \mathbb{S}^{d-1}$ est un vecteur unitaire échantillonné uniformément dans l'espace orthogonal à X_{i-1},
- $r_i \in [-1, 1]$ quantifie la distance entre X_{i-1} et X_i. r_i est distribué selon la densité de probabilité $f_{\mathcal{L}} : [-1, 1] \to [0, 1]$, appelée la *fonction latitude*.

En travaillant avec ce modèle, nos contributions présentées dans le chapitre 3 sont les suivantes. (i) Nous présentons un algorithme s'exécutant en temps polynomial basé sur un clustering hiérarchique pour estimer la fonction enveloppe $\mathbf{p}$. Nous fournissons des garanties théoriques pour notre algorithme lorsque le niveau de résolution optimal est connu. (ii) Nous proposons une procédure de sélection de

modèle basée sur l'heuristique de pente pour estimer un niveau de résolution réalisant un compromis biais/variance pertinent. (iii) Nous prouvons qu'il est possible d'estimer de façon consistante la matrice de Gram des positions latentes $G^* = n^{-1}\big(\langle X_i, X_j\rangle\big)_{i,j\in[n]}$ en norme Frobénius. (iv) Ce dernier résultat théorique motive l'approximation de la fonction latitude $f_{\mathcal{L}}$ en utilisant une estimateur par noyau à partir des approximations obtenues des distances latentes consécutives $\big(r_i\big)_{i\in\{2,\dots,n\}} = \big(\langle X_{i-1}, X_i\rangle\big)_{i\in\{2,\dots,n\}}$. (v) Nous prouvons que la connaissance des distances latentes est suffisante pour résoudre le problème de prédiction de liens. Ainsi, en se basant sur les estimateurs mentionnés ci-dessus de la fonction enveloppe $\mathbf{p}$, de la fonction latitude $f_{\mathcal{L}}$ et des distances latentes G^*, nous proposons une méthode pour estimer la probabilité de connexion entre les nœuds déjà présents dans le graphe et le nouvel arrivant. (vi) Nous fournissons une procédure de test pour déterminer si le graphe donné cache une dynamique markovienne latente non triviale ou si les noeuds ont été échantillonnés indépendamment et uniformément sur $\mathbb{S}^{d-1}$. (vii) Enfin, nous présentons les résultats de simulations numériques détaillées qui montrent les performances de nos méthodes.

L'analyse théorique des algorithmes utilisés pour réaliser des tâches d'estimation dans des modèles de graphes aléatoires dynamiques nécessite des outils probabilistes sophistiqués. Le modèle MRGG n'échappe pas à cette règle et la preuve d'un résultat essentiel du chapitre 3 a requis l'utilisation d'une inégalité de concentration pour une U-statistique dans un cadre dépendant. Les quelques articles existants dans la littérature qui abordaient ce problème difficile considéraient des hypothèses qui ne correspondaient pas à notre cadre. Cela nous a conduit à établir un nouveau résultat de concentration pour les U-statistiques d'une chaîne de Markov.

2. Outils probabilistes pour des variables aléatoires dépendantes et applications

2.1 Inégalité de concentration pour des U-statistiques dans un cadre dépendant (Chapitre 4)

Dans cette section, nous présentons un résultat de concentration pour des U-statistiques construites avec des variables aléatoires dépendantes. Ces travaux sont détaillés dans le chapitre 4 ainsi que dans l'article suivant.

> Duchemin et al. [2022b] Quentin Duchemin, Yohann De Castro, and Claire Lacour. Concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. *Bernoulli*, 2022b. URL https://hal.archives-ouvertes.fr/hal-03014763

Contexte et état de l'art. La concentration de la mesure est devenue omniprésente dans les communautés des Statistiques et du Machine Learning. En plus des graphes aléatoires (cf. Chapitres 2 et 3), on peut mentionner des applications de la concentration pour la sélection de modèles (voir Massart [2007] et Lerasle et al. [2016]), la théorie de l'apprentissage statistique (voir Clémençon et al. [2020]) ou l'apprentissage séquentiel (voir Wang et al. [2012]). Une contribution importante dans ce domaine est celle relative aux U-statistiques. Une U-statistique d'ordre m est une somme de la forme

$$\sum_{1\le i_1<\dots<i_m\le n} h_{i_1,\dots,i_m}(X_{i_1},\dots,X_{i_m}),$$

où $X_1,\dots,X_n$ sont des variables aléatoires prenant des valeurs dans un espace mesurable (E,Σ) et où $h_{i_1,\dots,i_m}$ sont des fonctions mesurables de m variables $h_{i_1,\dots,i_m} : E^m \to \mathbb{R}$.
Une inégalité exponentielle fondatrice pour les U-statistiques a été prouvée par Arcones and Giné [1993] à l'aide d'une approche de chaos. Leur résultat est valable pour des noyaux bornés et canoniques (ou dégénérés), ce qui signifie que pour tout $i_1,\dots,i_m \in [n] := \{1,\dots,n\}$ avec $i_1 < \dots < i_m$ et pout tout $x_1,\dots,x_m \in E$,

$$\big\|h_{i_1,\dots,i_m}\big\|_\infty < \infty \quad \text{et} \quad \forall j \in [1,n]\,,\ \mathbb{E}_{X_j}\Big[h_{i_1,\dots,i_m}(x_1,\dots,x_{j-1},X_j,x_{j+1},\dots,x_m)\Big] = 0\,.$$

Arcones et Giné ont prouvé que dans le cas dégénéré, les vitesses de convergence pour les U-statistiques sont de l'ordre de $n^{m/2}$. En s'appuyant sur des inégalités de moment de type Rosenthal, Giné et al. [2000] ont amélioré le résultat de Arcones and Giné [1993] en fournissant les quatre régimes optimaux de la queue de distribution, à savoir sous-gaussien, sous-exponentiel, et sous-Weibull d'ordres 2/3 et 1/2. Lorsque les noyaux ne sont pas bornés, certains résultats peuvent être étendus à condition que les variables aléatoires $h_{i_1,\dots,i_m}(X_{i_1},\dots,X_{i_m})$ aient des queues de distribution suffisamment légères [voir pour exemple Eichelsbacher and Schmock, 2003, Théorème 3.26]
Tous les résultats mentionnés ci-dessus considèrent que les variables aléatoires $(X_i)_{i\geq 1}$ sont indépendantes. Le comportement asymptotique des U-statistiques pour des variables aléatoires dépendantes a déjà été étudié par plusieurs articles [voir par exemple Bertail and Clémençon, 2011, Eichelsbacher and Schmock, 2001]. Toujours dans un cadre dépendant, les principaux travaux fournissant une inégalité de concentration pour les U-statistiques sont Borisov and Volodko [2015], Han [2018] et Shen et al. [2020]. Tous ces articles considèrent un noyau fixe (à savoir $h \equiv h_{i_1,\dots,i_m}$ pour tout $i_1,\dots,i_m$) défini sur $\mathbb{R}^d$ avec de fortes conditions de régularité. Pour la première fois, nous considérons dans cette thèse des fonctions dépendantes du temps, ce qui rend l'analyse théorique plus difficile puisque la méthode de *splitting* standard peut se révéler inutilisable (cf. section 4.2.5).

Hypothèses.
Nous considérons une chaîne de Markov $(X_i)_{i\geq 1}$ avec un noyau de transition $P : E \times E \to \mathbb{R}$ prenant des valeurs dans un espace mesurable (E, Σ), et nous introduisons des fonctions mesurables $h_{i,j} : E^2 \to \mathbb{R}$. Notre objectif est d'étudier les propriétés de concentration de la U-statistique d'ordre deux

$$U_{\text{stat}}(n) = \sum_{1\leq i<j\leq n} \left(h_{i,j}(X_i, X_j) - \mathbb{E}\left[h_{i,j}(X_i, X_j)\right]\right).$$

Nous travaillons sous les hypothèses suivantes.

1. **Ergodicité uniforme** *: La chaîne de Markov $(X_i)_{i\geq 1}$ est uniformément ergodique avec une mesure invariante π.*
2. **Noyau de transition borné** *: Il existe $\delta_M > 0$ et une mesure de probabilité ν telle que*
$$\forall x \in E,\ \forall A \in \Sigma, \quad P(x, A) \leq \delta_M \nu(A).$$
3. **Noyaux π-canoniques et bornés** *: Pour tout $i, j \in [n]$, $h_{i,j} : E \times E \to \mathbb{R}$ est mesurable, borné et π-canonique, c'est-à-dire*
$$\forall x, y \in E, \quad \mathbb{E}_\pi[h_{i,j}(X, x)] = \mathbb{E}_\pi[h_{i,j}(X, y)] = \mathbb{E}_\pi[h_{i,j}(x, X)] = \mathbb{E}_\pi[h_{i,j}(y, X)].$$
Cette espérance commune est notée $\mathbb{E}_\pi[h_{i,j}]$.
4. **Hypothèse technique** *: Au moins une des conditions suivantes est satisfaite*
   (i) Pour tout $i, j \in [n]$, $h_{i,j} \equiv h_{1,j}$, i.e. le noyau $h_{i,j}$ ne dépend pas de i.
   (ii) La distribution initiale de la chaîne est absolument continue par rapport à π et sa densité a un moment d'ordre p fini pour un certain $p \in (1, \infty]$.

Dans le chapitre 4, nous donnons des exemples de chaînes de Markov classiques vérifiant ces hypothèses.

Résultats. Pour la première fois, nous fournissons dans cette thèse une inégalité de concentration pour les U-statistiques d'ordre deux dans un cadre dépendant avec des noyaux qui peuvent dépendre des indices de la somme et qui ne sont pas supposés symétriques ou réguliers.
Nous prouvons dans un premier temps un résultat de concentration de type Hoeffding qui vaut sans aucune condition (ou sous une faible condition) sur la distribution initiale de la chaîne. En supposant que la chaîne de Markov $(X_i)_{i\geq 1}$ est stationnaire (ce qui signifie que X_1 est distribué selon π), nous prouvons une inégalité de concentration de type Bernstein qui conduit à une meilleure vitesse de convergence si les termes de variances sont petits. Nos principaux résultats sont résumés dans le Théorème 1.

Notre inégalité de concentration fait intervenir des quantités B_n et C_n qui peuvent être interprétées comme des écart-types et nous renvoyons au chapitre 4 pour leurs définitions précises. Afin de lire directement les termes dominants de notre inégalité de concentration à partir du Théorème 1, nous insistons sur le faut qu'il est toujours possible de borner grossièrement B_n et C_n comme suit

$$B_n \leq \sqrt{n}A \quad \text{and} \quad C_n \leq nA \quad \text{où} \quad A := 2 \max_{i,j} \|h_{i,j}\|_\infty.$$

Theorem 1

On considère que les hypothèses 1 à 4 sont satisfaites. Il existe alors des constantes $\beta, \kappa > 0$ telles que pour tout $u > 0$, on a avec probabilité au moins $1 - \beta e^{-u} \log n$,

$$U_{\text{stat}}(n) \leq \kappa \log(n) \Big(\left[C_n + A \log(n) \sqrt{n} \right] \sqrt{u} + \left[A + B_n \sqrt{n} \right] u + \left[2A\sqrt{n} \right] u^{3/2} + A \left[u^2 + \alpha_n \right] \Big),$$

où $\alpha_n = \begin{cases} \log(n) & \text{si la chaîne } (X_i)_{i \geq 1} \text{ est stationnaire} \\ n & \text{sinon} \end{cases}$.

Si l'hypothèse 4.(i) est vérifiée, le terme C_n peut être retiré de l'inégalité précédente.

Dans le chapitre 4, nous motivons l'utilisation de noyaux dépendant des indices de la somme en présentant deux exemples spécifiques empruntés aux domaines de la recherche d'information et des tests d'homogénéité. En considérant un cas particulier, nous montrons que notre inégalité de type Bernstein (obtenue lorsque $\alpha_n = \log(n)$ dans le Théorème 1) peut conduire à des vitesses de convergence significativement plus rapides.

Dans les trois sections suivantes, nous décrivons trois applications importantes du Théorème 1 en Statistiques et en Machine Learning. Ces contributions sont présentées dans le chapitre 5 et dans l'article suivant.

Duchemin et al. [2022a] Quentin Duchemin, Yohann De Castro, and Claire Lacour. Three rates of convergence or separation via U-statistics in a dependent framework. *JMLR*, 2022a. URL https://hal.archives-ouvertes.fr/hal-03603516

2.2 Estimation du spectre d'opérateurs intégraux signés via des algorithmes MCMC (Chapitre 5 Sec.5.3)

Contexte. Dans la théorie de l'apprentissage (comme par exemple pour l'Analyse en Composantes Principales), l'estimation des valeurs propres et/ou des vecteurs propres de matrices dépendantes des données est essentielle. Il apparaît que ces matrices peuvent souvent être interprétées comme les versions empiriques d'objets continus tels que les opérateurs intégraux. Comme souligné dans Rosasco et al. [2010], l'analyse théorique des algorithmes d'apprentissage mentionnés ci-dessus nécessite de quantifier la différence entre la structure propre des opérateurs empiriques et de leurs analogues continus. Dans cette thèse, nous étudions la convergence de la suite de spectres de matrices à noyau vers le spectre d'un opérateur intégral. Un travail fondateur dans ce domaine est celui de Adamczak and Bednorz [2015a] et, pour autant que nous le sachions, tous les résultats existants font l'hypothèse que le noyau est de type positif (*i.e.*, donne un opérateur intégral avec des valeurs propres positives). Pour la première fois, nous prouvons un résultat non asymptotique de convergence des spectres pour les noyaux qui ne sont pas de type positif. Nous prouvons également que les *algorithmes de Metropolis-Hastings indépendants* sont des schémas d'échantillonnage valides pour appliquer notre résultat.

Résultat. Nous considérons une chaîne de Markov $(X_n)_{n \geq 1}$ sur E satisfaisant les hypothèses 1 et 2 avec comme distribution de probabilité invariante π, et un certain noyau $h : E \times E \to \mathbb{R}$ satisfaisant les hypothèses suivantes.

> *$h : E \times E \to \mathbb{R}$ est une fonction bornée et symétrique de carré intégrable par rapport à la mesure $\pi \otimes \pi$. De plus, il existe des fonctions continues $\phi_r : E \to \mathbb{R}$, $r \in I$ (où $I = \mathbb{N}$ ou $I = 1, \ldots, N$) qui forment une base*

orthonormée de $L^2(\pi)$ et une suite de nombres réels $(\lambda_r)_{r\in I} \in \ell_2$ telles que l'on a ponctuellement

$$h(x,y) = \sum_{r\in I} \lambda_r \phi_r(x)\phi_r(y),$$

avec $\sup_{r\in I} \|\phi_r\|_\infty^2 < \infty$ *et* $\sup_{x\in E} \sum_{r\in I} |\lambda_r|\phi_r(x)^2 < \infty$.

On peut associer à h le noyau d'un opérateur linéaire $\mathbf{H}$ défini par

$$\mathbf{H}f(x) := \int_E h(x,y)f(y)d\pi(y).$$

Il s'agit d'un opérateur de Hilbert-Schmidt sur $L^2(\pi)$ et il possède donc un spectre réel constitué d'une suite de valeurs propres de carré sommable. Nous désignons les valeurs propres de $\mathbf{H}$ par $\lambda(\mathbf{H}) := (\lambda_1, \lambda_2, \dots)$. Pour un certain $n \in \mathbb{N}^*$, on considère $\mathbf{H}_n := \frac{1}{n}\left(h(X_i, X_j)\right)_{1\le i,j\le n}$ ayant pour valeurs propres $\lambda(\mathbf{H}_n)$.
Dans la section 5.3, nous prouvons que le spectre de $\mathbf{H}_n$ converge vers le spectre de l'opérateur intégral $\mathbf{H}$ lorsque $n \to \infty$. Plus précisément, il existe des constantes C, D telles que pour n suffisamment grand, on a avec probabilité au moins $1 - D/\sqrt{n}$,.

$$\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 \le \frac{C\log n}{\sqrt{n}} + 8 \sum_{i>\lceil n^{1/4}\rceil, i\in I} \lambda_i^2. \tag{1}$$

Signalons que le schéma de preuve de ce résultat généralise une approche déjà exploitée dans le chapitre 3.

Application. Considérons à présent un certain noyau h et une mesure de probabilité π satisfaisant les hypothèses précédentes. Notre objectif est de calculer les valeurs propres de l'opérateur intégral $\mathbf{H}$ associé à h. En pratique, π n'admet souvent pas d'expression sous forme close, une situation qui se présente typiquement dans un contexte bayésien où π est une certaine distribution *a posteriori*. Une approche standard pour résoudre ce problème consiste à s'appuyer sur des méthodes MCMC. Dans la section 5.3, nous adoptons cette stratégie et nous considérons le cas spécifique où E est un sous-ensemble borné de $\mathbb{R}^k$ équipé de tribu borélienne $\mathcal{B}(E)$. Nous considérons une densité de probabilité q sur E, appelée densité de proposition. Nous supposons que la mesure π sur E admet une densité f_π par rapport à la mesure de Lebesgue λ_{Leb} sur E et que les conditions suivantes sont satisfaites

$$\forall y \in E, \quad f_\pi(y), q(y) > 0 \quad \text{et} \quad \frac{q(y)}{f_\pi(y)} > \beta \ \text{ pour un certain } \beta > 0\,.$$

Dans ce contexte, nous prouvons dans la section 5.3 qu'une chaîne de Markov $(X_i)_{i\ge1}$ obtenue à partir d'un algorithme de Metropolis-Hastings indépendant avec une loi de proposition $q\lambda_{Leb}$ satisfait les hypothèses 1 et 2. Nous déduisons que l'on peut estimer les valeurs propres de $\mathbf{H}$ en calculant celles de $\mathbf{H}_n$, et l'équation (1.1) quantifie la distance entre les deux spectres.

2.3 Bornes de généralisation pour l'apprentissage séquentiel avec fonction de perte bivariée (Chapitre 5 Sec.5.4)

Contexte. En Machine Learning, les algorithmes batch accumulent les données sur une période et n'entraînent les modèles qu'une fois le processus d'acquisition des données terminé. L'apprentissage de type batch présente certaines limites, notamment lorsque $i)$ les données arrivent au cours du temps (par exemple, les cours de la bourse) et que nous devons nous adapter rapidement aux changements, ou $ii)$ pour les problèmes d'apprentissage à grande échelle où traiter d'un seul bloc l'ensemble des données peut se réveler trop coûteux d'un point de vue computationnel. Les algorithmes séquentiels ont été conçus pour résoudre efficacement les problèmes d'apprentissage dans de telles situations : ils traitent les données à la volée et tentent d'améliorer le modèle appris au fil du temps en fonction des nouvelles observations. Une façon d'analyser la performance des algorithmes d'apprentissage séquentiels est de considérer la notion de *regret* qui compare les pertes induites par les décisions prises par l'algorithme au cours du temps et la perte qui aurait été subie en prenant une décision optimale. Au

cours de la dernière décennie, les chercheurs ne se sont pas seulement intéressés à la notion de regret mais ont examiné les algorithmes séquentiels sous un angle différent en se demandant quelles seraient leurs performances sur de futures données. Cette question n'a de sens que si l'on suppose que la suite d'observations provient d'un certain processus stochastique. Comme exprimé dans Agarwal and Duchi [2012], *"si la suite d'exemples est générée par un processus stochastique, l'algorithme d'apprentissage séquentiel peut-il fournir un bon prédicteur pour les échantillons futurs issus du même processus ?"*
Les performances de généralisation des algorithmes séquentiels avec des fonctions de perte univariées ont déjà été largement étudiées pour des observations aussi bien i.i.d. que dépendantes. En ce qui concerne les problèmes d'apprentissage séquentiels avec une fonction de perte bivariée, des bornes de généralisation ont été obtenues pour des données i.i.d. mais peu de travaux considèrent le cas d'observations dépendantes. Cette thèse fournit l'un des premiers résultats dans ce cadre difficile. La Figure 2 positionne précisément notre travail dans la littérature.

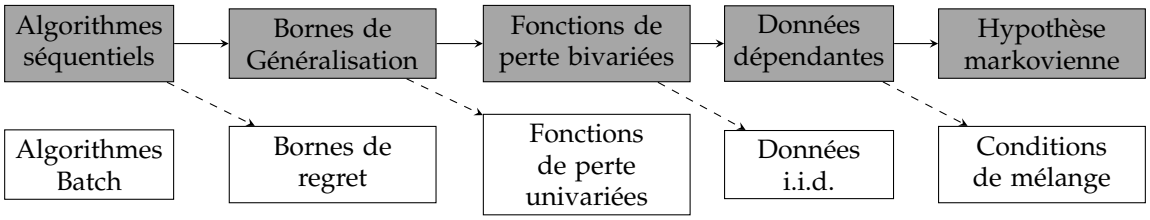


FIGURE 2 : Positionnement de nos contributions dans la littérature existante pour l'analyse des algorithmes séquentiels.

Résultats. Inspirés par le problème de *ranking*, nous considérons une fonction $f : E \to \mathbb{R}$ qui définit l'ordre des objets dans E. Nous cherchons à trouver une approximation pertinente de l'ordre des éléments de E en sélectionnant une fonction h (appelée fonction *hypothèse*) dans un espace $\mathcal{H}$ en se basant sur l'observation de la suite aléatoire $(X_i, f(X_i))_{1\leq i\leq n}$ où $(X_i)_{i\geq 1}$ est une chaîne de Markov réversible satisfaisant les hypothèses 1 et 2. Pour mesurer la performance d'une hypothèse donnée $h : E \times E \to \mathbb{R}$, nous utilisons une fonction de perte bivariée de la forme $\ell(h, X, U)$. Typiquement, on peut utiliser la fonction de perte nommée *misranking loss* et définie par

$$\ell(h, x, u) = \mathbb{1}_{\{(f(x)-f(u))h(x,u)<0\}},$$

qui vaut 1 si les exemples sont classés dans le mauvais ordre et 0 sinon. Le but du problème d'apprentissage est de trouver une hypothèse h qui minimise l'espérance du *misranking risk*.

$$\mathcal{R}(h) := \mathbb{E}_{(X,X')\sim\pi\otimes\pi}\big[\ell(h, X, X')\big].$$

Dans le contexte de l'apprentissage séquentiel, à chaque pas de temps t l'algorithme choisit une certaine hypothèse $h_t \in \mathcal{H}$ en observant uniquement la suite $(X_i, f(X_i))_{i\leq t}$ jusqu'au temps t. Nous travaillons avec les hypothèses suivantes.

- $(\mathcal{H}, \|\cdot\|_\infty)$ *est compact et satisfait*

  $$\log \mathcal{N}(\mathcal{H}, \eta) = \mathcal{O}(\eta^{-\theta}),$$

  où $\mathcal{N}(\mathcal{H}, \eta)$ *est le* L^∞*-covering number de* $\mathcal{H}$ *et où* $\theta > 0$.
- *La fonction de perte* $\ell : \mathcal{H} \times E \to [0, 1]$ *est telle que*

  $$\ell(h, x_1, x_2) = \phi(f(x_1) - f(x_2), h(x_1, x_2)),$$

  où $\phi : \mathbb{R} \times \mathbb{R} \to [0, 1]$ *est Lipschitz par rapport à la seconde coordonnée.*

Nos contributions présentées dans la section 5.4 sont les suivantes :

1. Nous introduisons un nouveau *risque empirique*, désigné par $\mathcal{M}^n := \mathcal{M}^n(h_1, \ldots, h_{n-1-b_n})$. Ce dernier dépend de la quantité clé b_n qui peut être interprétée comme un facteur d'oubli. b_n est de

l'ordre de $\log n$ et sa définition implique une constante qui rend compte des propriétés de mélange de la chaîne de Markov.

2. Nous donnons des bornes d'erreur non asymptotiques entre $\mathcal{M}^n$ et le vrai risque moyen associé à la suite d'hypothèses $(h_t)_{t\in[n]}$ générée par l'algorithme séquentiel considéré. Plus précisément, notant $c_n = \lfloor cn \rfloor$ pour un certain $c \in (0,1)$, nous montrons que

$$\left| \frac{1}{n-c_n} \sum_{t=c_n}^{n-1} \mathcal{R}(h_{t-b_n}) - \mathcal{M}^n \right| = \mathcal{O}_{\mathbb{P}}\left[\frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}} \right].$$

3. Nous convertissons une borne de regret d'un algorithme séquentiel donné en un contrôle de l'excès de risque. Ce type de résultat est connu dans la littérature sous le nom de "conversion *online-to-batch*". Plus précisément, en considérant un algorithme qui atteint une borne de regret $\mathfrak{R}_n$ c'est-à-dire générant une suite d'hypothèses $(h_t)_{t\in[n]}$ telle que

$$\mathcal{M}^n \leq \inf_{h\in\mathcal{H}} \left\{ \mathcal{M}^n(h,\dots,h) \right\} + \mathfrak{R}_n,$$

nous montrons que le risque moyen de l'ensemble des hypothèses $(h_t)_{t\in[n]}$ satisfait à

$$\frac{1}{n-c_n} \sum_{t=c_n}^{n-1} \mathcal{R}(h_{t-b_n}) - \min_{h\in\mathcal{H}} \mathcal{R}(h) = \mathcal{O}_{\mathbb{P}}\left[\frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}} + \mathfrak{R}_n \right].$$

4. Nous donnons une procédure de sélection d'hypothèse $\widehat{h}$ parmi l'ensemble $\{h_t,\ t \in [n]\}$ réalisant ce risque moyen.

2.4 Test d'adéquation adaptatif pour la densité de la mesure invariante d'une chaine de Markov (Chapitre 5 Sec.5.5)

Contexte. Plusieurs travaux ont déjà proposé des tests d'adéquation pour la densité de la distribution invariante d'une suite de variables aléatoires dépendantes et nous pouvons citer par exemple Bai [2003], Chwialkowski et al. [2016], Li and Tkacz [2001]. Dans tous les articles mentionnés ci-dessus, certaines propriétés asymptotiques de la statistique de test sont décrites mais aucune analyse non-asymptotique des méthodes n'est menée. Pour autant que nous le sachions, nous fournissons pour la première fois une condition non asymptotique sur les classes d'alternatives garantissant que le test statistique atteint une puissance prescrite dans un cadre dépendant.

Résultat. Nous considérons une chaîne de Markov $X_1,\dots,X_n$ avec une distribution invariante π admettant une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$ et satisfaisant les hypothèses 1 et 2. On considère f_0 une densité donnée dans $L^2(\lambda_{Leb})$, $\alpha \in]0,1[$, et nous supposons que f appartient à $L^2(\lambda_{Leb})$. Sous ces hypothèses, nous construisons un test de niveau α pour l'hypothèse nulle "$f = f_0$" contre l'alternative "$f \neq f_0$" à partir de l'observation $(X_1,\dots,X_n)$. Le test est basé sur l'estimation de $\|f-f_0\|_2^2$ qui s'écrit encore comme $\|f\|_2^2 + \|f_0\|_2^2 - 2\langle f, f_0\rangle$. $\langle f, f_0\rangle$ est généralement estimé par une approche empirique $\sum_{i=1}^n f_0(X_i)/n$ et la pierre angulaire de notre approche consiste à estimer $\|f\|_2^2$. Nous nous appuyons sur les travaux de Fromont and Laurent [2006] et nous introduisons un ensemble $\{S_m, m \in \mathcal{M}\}$ de sous-espaces vectoriels de $L^2(\lambda_{Leb})$. Dans la section 5.5, nous considérons trois collections différentes de sous-espaces vectoriels $\{S_m, m \in \mathcal{M}\}$, à savoir les fonctions constantes par morceaux, les ondelettes et les polynômes trigonométriques. Pour tout m dans $\mathcal{M}$, nous considérons $\{p_l, l \in \mathcal{L}_m\}$ une base orthonormée de S_m. Notant Π_{S_m} la projection orthogonale sur S_m, nous considérons l'estimateur de $\|\Pi_{S_m}(f)\|_2^2$ donné par

$$\widehat{\theta}_m = \frac{1}{n(n-1)} \sum_{l\in\mathcal{L}_m} \sum_{i\neq j=1}^{n} p_l(X_i)p_l(X_j).$$

La distance $\|f - f_0\|_2^2$ peut alors être estimée par

$$\widehat{T}_m = \widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} f_0(X_i),$$

pour tout m dans $\mathcal{M}$. En notant $t_m(u)$ le quantile d'ordre $(1 - u_\alpha)$ de la loi de $\widehat{T}_m$ sous l'hypothèse nulle "$f = f_0$" et en considérant

$$u_\alpha = \sup_{u\in]0,1[} \mathbb{P}_{f_0}\left(\sup_{m\in\mathcal{M}} (\widehat{T}_m - t_m(u)) > 0\right) \leq \alpha,$$

nous introduisons la statistique de test T_α définie par

$$T_\alpha = \sup_{m\in\mathcal{M}} (\widehat{T}_m - t_m(u_\alpha)).$$

Le test consiste à rejeter l'hypothèse nulle si T_α est positif. Cette approche peut être interprétée comme une procédure de tests multiples. En effet, pour chaque m dans $\mathcal{M}$, on construit un test de niveau u_α pour l'hypothèse nulle "$f = f_0$" en rejetant cette hypothèse si $\widehat{T}_m$ est plus grand que son quantile d'ordre $(1-u_\alpha)$ sous "$f = f_0$". On obtient donc une collection de tests et on décide de rejeter l'hypothèse nulle si pour au moins un test de la collection l'hypothèse "$f = f_0$" est rejetée.
Dans la section 5.5, nous fournissons une borne supérieure sur la *vitesse de séparation* pour des classes spécifiques d'alternatives qui incluent certains espaces de Besov. Rappelons que la vitesse de séparation associée à une classe d'alternatives $\mathcal{B}$ et à un certain $\gamma \in (0, 1)$ est définie comme le plus petit réel $\rho > 0$ tel que pour tout $f_1 \in \mathcal{B}$ avec $\|f_0 - f_1\|_2 > \rho$, la puissance du test pour l'alternative "$f = f_1$" est supérieure à $1 - \gamma$. Dans la section 5.5, nous prouvons que la vitesse de séparation de notre méthode est majorée par

$$\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}},$$

où le paramètre $s > 0$ quantifie la régularité des densités appartenant à la classe d'alternatives considérée. Nous comparons ce résultat à la vitesse minimax connue pour un test adaptatif dans le cadre de données i.i.d. qui est de l'ordre de $(\sqrt{\log\log n}/n)^{2s/(4s+1)}$ (voir Ingster [1993]). À la fin de la section 5.5, nous présentons également les résultats de simulations comparant notre approche au test de Kolmogorov-Smirnov et au test du χ^2 pour différentes chaînes de Markov.

Nous revenons maintenant à l'une des motivations initiales de cette thèse : le problème de prédiction de liens dans les graphes aléatoires. Dans les chapitres 2 et 3, nous avons abordé cette question dans des modèles de graphon où les noeuds rejoignent le graphe au cours du temps avec une représentation latente dépendant du dernier entrant. Nous supposerons dans la section suivante que les représentations des nœuds $X_i \in \mathbb{R}^d$ sont observées et que le graphon appartient à une classe paramétrique de grande dimension basée sur la régression logistique. De nombreux problèmes en Machine Learning se place dans le cadre de la grande dimension qui rend souvent la tâche d'estimation mal posée. Pour faire face à ce problème, une approche classique consiste à $i)$ supposer une certaine structure sur le signal cible (typiquement une hypothèse de parcimonie), $ii)$ effectuer une étape de sélection de modèle et $iii)$ estimer ensuite le signal d'intérêt en utilisant le modèle sélectionné. Puisque les données ont été exploitées pour sélectionner le modèle, l'utilisation des méthodes standard d'inférence peut conduire à des propriétés statistiques indésirables. L'inférence post-sélection (PSI) vise à résoudre ce problème en prenant en compte l'événement de sélection pour fournir des procédures d'inférence valides. Dans le chapitre 6, nous poussons plus loin l'état actuel des connaissances sur les méthodes de PSI dans les modèles linéaires généralisés en nous concentrant tout particulièrement sur la régression logistique parcimonieuse.

3. Inférence sélective pour le régression logistique parcimonieuse (Chapitre 6)

Motivations. Toujours motivés par les problèmes de prédiction de liens dans les graphes aléatoires, nous considérons maintenant un cadre où pour chaque nœud $i \in [n]$ d'un graphe simple et non dirigé de taille n, nous observons un vecteur de caractéristiques $X_i \in \mathbb{R}^d$. Étant donné les observations X_i, X_j sur les nœuds i et j du réseau, nous considérons que les nœuds i et j sont connectés avec une probabilité $W(X_i, X_j)$ pour une certaine fonction symétrique $W : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$. Inspiré par le travail de Berthet and Baldin [2020], nous considérons que le graphon W appartient à une classe de grande dimension basée sur la régression logistique. Notant $\sigma : x \mapsto (1+\exp(-x))^{-1}$ la fonction sigmoïde, nous supposons qu'il existe une certaine matrice $\Theta^* \in \mathbb{R}^{d\times d}$ tel que $W(X_i, X_j) = \sigma(X_i^\top \Theta^* X_j)$. Notre objectif est d'estimer la matrice inconnue Θ^* à partir de l'observation de la matrice d'adjacence du graphe A, et des variables explicatives connues $\mathbb{X} := [X_1 \ \dots \ X_n] \in \mathbb{R}^{d\times n}$, et de mener des procédures d'inférence (*i.e.* construire des tests d'hypothèse ou des intervalles de confiance). Nous faisons l'hypothèse que Θ^* est parcimonieux, ce qui signifie que seul un petit sous-ensemble des d covariables observées influence effectivement la connexion entre deux nœuds du graphe. Dans cette situation, l'approche classique adoptée par les statisticiens consiste à suivre un protocole en trois étapes.

- *Sélection de modèle* : Sur la base des données observées, le statisticien sélectionne un sous-ensemble d'entrées actives dans la matrice Θ. Une approche standard consiste à calculer l'estimateur du maximum de vraisemblance avec une pénalité ℓ_1

$$\widehat{\Theta}^\lambda \in \arg\min_{\Theta\in\mathbb{R}^{d\times d}} \big\{ -\log \mathbb{P}_\Theta(A \,|\, \mathbb{X}) + \lambda\|\Theta\|_1 \big\}, \tag{2}$$

  qui peut être réécrite comme une régression logistique classique. En effet, en notant $\mathrm{vec}(B) \in \mathbb{R}^{p^2}$ la forme vectorisée d'une matrice $B \in \mathbb{R}^{p\times p}$, on a

$$X_j^\top \Theta^* X_i = \mathrm{Tr}(X_i X_j^\top \Theta^*) = \langle \mathrm{vec}(X_i X_j^\top), \mathrm{vec}(\Theta^*) \rangle.$$

  Ainsi, en adoptant des notations évidentes, le problème d'optimisation (2) est équivalent à

$$\widehat{\vartheta}^\lambda \in \arg\min_{\vartheta\in\mathbb{R}^{d^2}} \big\{ -\log \mathbb{P}_\vartheta(Y \,|\, \mathbf{X}) + \lambda\|\vartheta\|_1 \big\},$$

  où $Y = \mathrm{vec}(A)$, $\mathbf{X} = [\mathrm{vec}(X_1X_1^\top)\ \mathrm{vec}(X_1X_2^\top)\ \dots\ \mathrm{vec}(X_nX_n^\top)] \in \mathbb{R}^{d^2\times n^2}$ et où $\widehat{\vartheta}^\lambda = \mathrm{vec}(\widehat{\Theta}^\lambda)$. Nous définissons alors l'ensemble des entrées actives par $M := \widehat{M}(Y) := \{i \,|\, \widehat{\vartheta}^\lambda_i \neq 0\}$.

- *Estimation* : Le statisticien calcule l'estimateur du maximum de vraisemblance (MLE) en utilisant uniquement les variables dans M

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^{|M|}} \big\{ -\log \mathbb{P}_{\vartheta(\theta)}(Y \,|\, \mathbf{X}) \big\},$$

  où pour tout $\theta \in \mathbb{R}^{|M|}$, $\vartheta(\theta) \in \mathbb{R}^{d^2}$ est tel que $\vartheta_{-M}(\theta) = 0$ et $\vartheta_M(\theta) = \theta$.

- *Inférence* : Le statisticien effectue des tests d'hypothèse et fournit des intervalles de confiance.

Inférence Post-Sélection (PSI). L'étape de sélection de modèle nécessite le choix de l'hyperparamètre λ, ce qui est effectué en pratique en utilisant les données. Dans ce contexte, l'application de méthodes d'inférence standard sans tenir compte de l'utilisation des données pour sélectionner le modèle conduit généralement à des propriétés statistiques indésirables (cf. Pötscher [1991]). L'inférence post-sélection vise à résoudre ce problème. Il s'agit de construire des procédures d'inférence en considérant que le vecteur d'observations Y est distribué selon la loi $\mathbb{P}_{\vartheta^*}(\tilde{Y} \,|\, \mathbf{X}, \{\tilde{Y} \in E_M\})$. Dans cette distribution conditionnelle, $E_M := \{\tilde{Y} \,|\, M = \widehat{M}(\tilde{Y})\}$ est appelé l'*événement de sélection* et correspond à l'ensemble de tous les graphes aléatoires ayant pour matrice d'adjacence vectorisée A (telle que $\mathrm{vec}(A) = \tilde{Y}$) qui auraient conduit au même ensemble de variables actives que le graphe avec matrice d'adjacence B telle que $\mathrm{vec}(B) = Y$. La PSI dans le contexte de la régression linéaire a connu un intérêt grandissant ces dernières années, en particulier grâce à l'importante contribution de Lee et al. [2016]. Dans ce dernier article, les auteurs prouvent que dans le modèle linéaire avec un bruit gaussien, la distribution de la

variable de réponse conditionnellement à l'événement de sélection est un mélange de gaussiennes multivariées tronquées. Ce résultat est une conséquence du lemme dit *polyhédral* et permet de fournir des procédures exactes de PSI. Des méthodes de PSI en dehors du modèle linéaire avec bruit gaussien ont été étudiées récemment et on peut citer par exemple Fithian et al. [2014], Taylor and Tibshirani [2018], Tian and Taylor [2017], Tian et al. [2018], Tibshirani et al. [2018]. Malgré l'omniprésence du modèle de la régression logistique dans les applications, il reste l'un des cadres où les méthodes de PSI existantes pour les modèles linéaires généralisés (GLMs) sont soit inadaptées [cf. Fithian et al., 2014, Section 6.3], soit manquent de garanties théoriques [cf. Taylor and Tibshirani, 2018].

Contributions. Dans le chapitre 6, (i) nous donnons une nouvelle formulation de l'événement de sélection dans les GLMs mettant en lumière le difféomorphisme essentiel Ψ qui porte l'information géométrique du problème, (ii) nous fournissons une nouvelle perspective sur l'inférence post-sélection dans les GLMs à travers l'approche de MLE conditionnel dont Ψ est un ingrédient clé, (iii) nous introduisons des conditions suffisantes dans les GLMs pour obtenir des procédures de PSI asymptotiquement valides basées sur l'approche de MLE conditionnel. Par la suite, nous nous concentrons sur le cas spécifique de la régression logistique : (iv) Nous prouvons - sous certaines hypothèses - que les conditions suffisantes de (iii) sont satisfaites pour le modèle logistique. (v) Cela nous permet de donner des procédures de PSI asymptotiquement valides pour la régression logistique et nous appuyons nos résultats théoriques par des simulations. (vi) Enfin, nous présentons une comparaison approfondie entre notre travail et l'heuristique de Taylor and Tibshirani [2018] qui est actuellement considérée comme la meilleure à utiliser dans le contexte de la régression logistique parcimonieuse [cf. Fithian et al., 2014, Section 6.3].
Ce travail sur les méthodes d'inférence post-sélection dans le GLMs est présenté dans le chapitre 6 et correspond à l'article suivant.

Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. A new procedure for Selective Inference with the Generalized Linear Lasso. 2022. URL https://hal.archives-ouvertes.fr/hal-03622196

Chapter 1

Introduction

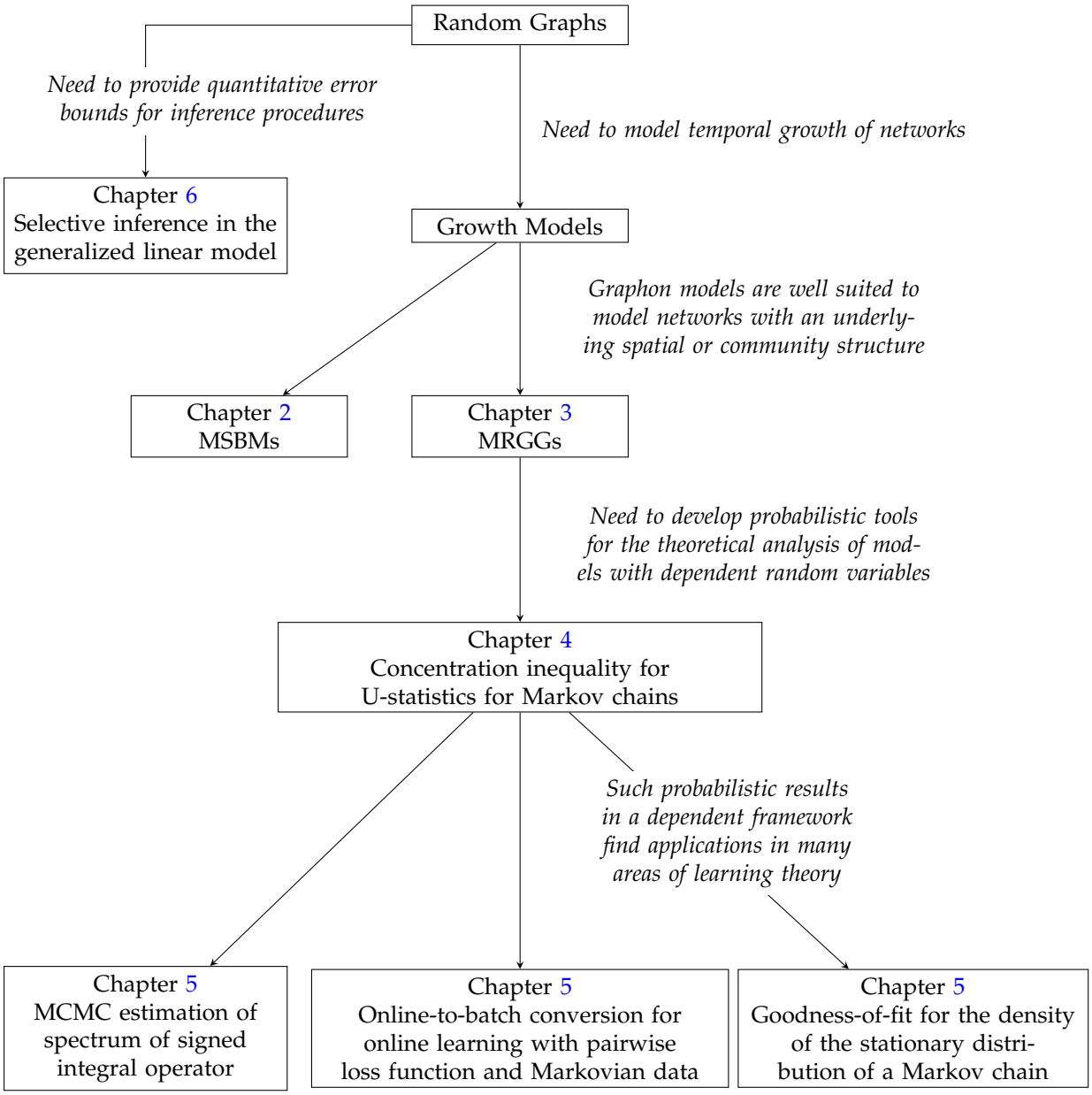


Figure 1.1: Structure of the thesis.

1.1 Growth models for random graphs

1.1.1 Networks modeling through random graph

Random graph modeling. Graphs are nowadays widely used in applications to model real world complex systems. Since they are high dimensional objects, one needs to assume some structure on the data of interest to be able to efficiently extract information on the studied system. To this purpose, a large number of models of random graphs have been already introduced. The most simple one is the Erdös-Renyi model $G(n,p)$ in which each edge between pairs of n nodes is present in the graph with some probability $p \in (0,1)$. One can also mention the scale-free network model of Barabasi and Albert [Barabási, 2009] or the small-world networks of Watts and Strogatz [Watts and Strogatz, 1998]. We refer to Channarond [2015] for an introduction to the most famous random graph models. On real world problems, it appears that there often exist some relevant variables accounting for the heterogeneity of the observations. Most of the time, these explanatory variables are unknown and carry a precious information on the studied system. To deal with such cases, latent space models for network data emerged (see Smith et al. [2019]). Ones of the most studied latent models are the *community based random graphs* where each node is assumed to belong to one (or multiple) community while the connection probabilities between two nodes in the graph depend on their respective membership. The well-known Stochastic Block Model (SBM) has received increasing attention in the recent years and we refer to Abbe [2017] for a nice introduction to this model and the statistical and algorithmic questions at stake. In the previous mentioned latent space models the intrinsic geometry of the problem is not taken into account. However, it is known that the underlying spatial structure of network is an important property since geometry drastically affects the topology of networks (see Barthélemy [2011] and Smith et al. [2019]). To deal with embedded complex systems, spatial random graph models have been studied such as the Random Geometric Graph (RGG).

Graphon models: the particular examples of SBMs and RGGs. Both the SBM and the RGG can be understood as specific examples of graphon models [cf. Lovász, 2012]. In a graphon model, we consider some latent space $\mathcal{X}$ and a symmetric kernel function $W : \mathcal{X} \times \mathcal{X} \to [0,1]$. To build a non-oriented and simple graph of size n from the graphon model associated to W, one first needs to sample latent positions $(X_i)_{i\in[n]} \in \mathcal{X}^n$. Two vertices $i, j \in [n]$ with $i \neq j$ are then connected with probability $W(X_i, X_j)$.
SBMs and RGGs are distinguished by their different topological structure. This is the reason why they are used in practice to model physical phenomena of different natures. Taking the example of social networks, we can consider as a first approximation [as explained in Péché and Perchet, 2020] that a connection between two users can occur for two main reasons: (i) either they are childhood friends (meaning that users are endogenously similar with close geometric latent representations) or (ii) they share the same political views (meaning that users are exogenously similar and belong to the same community). Hence both models are of interest for modern applications and have received an increasing interest in the past decades. SBMs and RGGs are also of independent interest for the interesting mathematical questions they raise.

Time prediction. In practice, real-world networks are evolving through time and we can mention the example of spread of diseases [cf. Barthélemy, 2011, Section 5.6.3]. To bridge the gap between random graph models and the complexity of real data, stochastic models for network evolution have been extensively studied in the recent years. Most of them rely on the latent space approach where the size of the graph is fixed and edges and/or latent representations can change along time [cf. Lei and Rinaldo, 2015, Matias and Miele, 2015, 2017, Xie and Rogers, 2016, Xie et al., 2015, Xu and Hero, 2014]. In this thesis, we adopt a different perspective. Motivated by link prediction problems, we will focus on *growth models*, namely random graph models in which a node is added at each new time step in the network and is connected to other vertices in the graph according to some probabilistic rule that needs to be specified. In the last decade, growth models for random graphs with a spatial structure have gained an increased interest. One can mention Jordan and Wade [2015], Papadopoulos et al. [2012] and Zuev et al. [2015] where geometric variants of the preferential attachment model are introduced with one new node entering the graph at each time step.
In the following, we propose growth models based on the SBM and on the RGG where the latent node attributes are sampled using a Markovian dynamic and we solve link prediction tasks by estimating the probability of connection with nodes already present in the graph and future comers.

1.1.2 A growth model for SBMs (Chapter 2)

Brief presentation of the Stochastic Block Model. The SBM is a perfect playground to study the existence of phase transition phenomena and information-theoretic/computational thresholds. In the SBM, we observe the adjacency matrix of the graph and we aim at obtaining information on the hidden latent communities. Different recovery requirements can be studied in the SBMs such as *exact recovery* (where one aims at recovering correctly the entire partition with high probability) or *weak recovery* (where the algorithm should provide with high probability a positively correlated partition). A large span of methods have been developed to solve these problems supported with theoretical guarantees. One can mention two-round algorithms via graph-splitting, semi-definite programming, or linearized belief propagation just to name of few. Nevertheless, with its original formulation, the SBM copes with several issues regarding their practical usage for applications.

- ✓ *The sparse regime.* When the average degree of nodes in the graph is constant (i.e., do not depend on n), SBMs are not connected with high probability and exact recovery cannot be achieved. In this case, the goal is to find an algorithm that solves the so-called weak recovery (or detection) problem, meaning that it divides the graph's vertices into two sets such that vertices from two different communities have different probabilities of being assigned to one of the sets. Most of methods used to solve the exact recovery problem - such spectral methods on Laplacian matrices of the graph - fail to solve detection in the sparse regime. In Krzakala et al. [2013], the authors introduced a new graph representation matrix named the non-backtracking matrix B and claimed that a spectral method on B could solve detection in the sparse regime. This *spectral redemption* through the non-backtracking operator was rigorously proved by Bordenave et al. [2018] for the symmetric SBM with two communities. The general conjecture for arbitrary many symmetric or asymmetric communities is settled later in Abbe and Sandon [2015b] relying on a higher-order nonbacktracking operator and a message passing implementation.
- ✓ *The degree heterogeneity.* Another main limit of spectral methods for real application is when the studied network present degree heterogeneity. Working with the degree correlated SBM, Dall'Amico and Couillet [2019] proved that the Bethe Hessian matrix can be used to solve community detection in sparse graphs with inhomogeneous degrees.
- ✓ *Partial recovery bounds.* Interpolating between the exact and weak recovery requirements, one main challenge in SBMs is to understand the inherent connection between an appropriate signal to noise ratio (SNR) and the agreement, namely the proportion of correct alignment between the community allocation found by the algorithm and the ground truth labels. In Giraud and Verzelen [2019], the authors proposed a Semi-Definite Program (SDP) algorithm followed by a rounding step to estimate communities in a graph and proved that the agreement goes exponentially fast to 1.

The previous discussion was focused on static graph. However, in many applications, one has access to multiple snapshots of the same graph that evolves along time. This is the case of networks representing physical proximity of mobile agents or biological and chemical evolution of group members and we refer to Holme [2015] for a review. In order to extract temporal information on the system of interest, several works have extended the SBM to model the dynamic structure of the studied networks. In Matias and Miele [2015], a variant of the SBM is considered where the temporal evolution is modeled through a discrete hidden Markov chain on the nodes membership and where the connection probabilities also evolve through time. Following the work of Karrer and Newman [2011], Lei and Rinaldo [2015] study the Degree Corrected SBM where the degree of the nodes can vary within the same community. We can also mention Keriven and Vaiter [2022] or Dall'Amico et al. [2020].
The above mentioned works are mainly considering SBMs where membership of nodes or edges can evolve with time, but only few papers are interested in growth model for SBMs (meaning that the size of the graph increases along time) and we aim at filling this gap.

The Markov Stochastic Block Model. In Chapter 2, we introduce the Markov SBM (MSBM): an extension of the Stochastic Block Model where communities of the nodes are assigned through a Markovian dynamic. We want to provide efficient and reliable methods to solve

- link prediction problems where we aim at computing the probability of connection between nodes in the graph and some future entrant,

- or collaborative filtering tasks where we want to infer the hidden community of some node if we have only partial information on how this node connects to the rest of the graph.

In the MSBM, we consider that at each time step i, a new node is entering the network with a latent community $C_i \in [K]$ (for some fixed positive integer K) that is sampled from the probability distribution $P_{C_i,\cdot}$ where $P \in [0,1]^{K\times K}$ is the transition matrix of the positive recurrent Markov chain $(C_i)_{i\in[n]}$. Once the community of each node is assigned, we draw an edge between the nodes i and j with probability Q_{C_i,C_j} where $Q \in [0,1]^{K\times K}$ is the connectivity matrix. In this model, one can use standard methods to recover the hidden communities of the nodes from the observed adjacency matrix of the graph and thus provide estimates of the model parameters P and Q. In order to solve link prediction problems or collaborative filtering tasks, a natural approach is to rely on plug-in methods using the above mentioned estimates of P, Q and $(C_i)_{i\in[n]}$. Nevertheless, we show in Chapter 2 that this plug-in procedure can lead to large estimation errors because it is highly sensitive to clustering errors made by the algorithm. In Chapter 2, we propose a general methodology to solve link prediction and collaborative filtering tasks in the MSBM that is adaptive to local clustering errors and that is shown to be much more reliable for applications. Our contributions are the following.

- General methods for reliable time prediction in MSBMs.
  $i)$ Establishing a connection between MSBMs and Hidden Markov Models, we propose to learn the so-called *emission probabilities* that correspond for any $c, \hat{c} \in [K]$ to the probability for some node with community $c \in [K]$ to be assigned to community $\hat{c}$ by the clustering algorithm considered. These quantities allow us to design reliable link prediction and collaborative filtering methods that can account for local errors in the estimates of the hidden communities.
  $ii)$ We also show how the learned emission probabilities can be used for model selection, *i.e.* to estimate the unknown number of latent communities K.
  Let us stress that these methods can be used with any clustering algorithm. In order to conduct numerical experiments and to provide some theoretical guarantees, we work in Chapter 2 with the recent Semi-Definite Programming (SDP) algorithm from Giraud and Verzelen [2019].

- Theoretical guarantees.
  $iii)$ We identify a relevant signal-to-noise ratio (SNR) in our framework and we prove that the partition of nodes obtained from the SDP algorithm leads to a misclassifaction error that decays exponentially fast with respect to this SNR.
  $iv)$ We give estimates of the parameters of the MSBMs that are proved to be consistent.

- Numerical aspects.
  $v)$ As far as we know, we are the first to provide an implementation of the SDP method from Giraud and Verzelen [2019]. This work has required to code a tricky rounding step from Charikar et al. [2002].
  $vi)$ We provide extensive numerical results of our methods on both simulated and real data.

This work on MSBMs presented in Chapter 2 can also be found in the following paper.

> Duchemin [2022] Quentin Duchemin. Reliable Time Prediction in the Markov Stochastic Block Model. preprint, March 2022. URL https://hal.archives-ouvertes.fr/hal-02536727

1.1.3 A growth model for RGGs (Chapter 3)

Brief presentation of Random Geometric Graphs. The RGG model was first introduced by Gilbert [1961] to model the communications between radio stations. Gilbert's original model was defined as follows: pick points in $\mathbb{R}^2$ according to a Poisson Point Process of intensity one and join two if their distance is less than some parameter $r > 0$. The Random Geometric Graph model was extended to other latent spaces such as the hypercube $[0,1]^d$, the Euclidean sphere or compact Lie group Méliot [2019]. A large body of literature has been devoted to studying the properties of low-dimensional Random Geometric Graphs Penrose [2003], Dall and Christensen [2002], Bollobás [2001]. RGGs have found applications in a very large span of fields. One can mention wireless networks Haenggi et al. [2009], gossip algorithms Wang and Lin [2014], spread of a virus Preciado and Jadbabaie [2009], protein-protein interactions Higham et al. [2008a]. The ubiquity of this random graph model to faithfully represent real world networks has motivated a great interest for its theoretical study.

To quote Bollobás [2001], *"One of the main aims of the theory of random graphs is to determine when a given property is likely to appear."* In this direction, several works tried to identify structure in networks through testing procedure, see for example Ghoshdastidar et al. [2020]. Regarding RGGs, most of the results have been established in the low dimensional regime $d \leq 3$ [cf. Barthélemy, 2011, Penrose, 2003]. However, applications of RGGs to cluster analysis and the interest in the statistics of high-dimensional data sets have motivated the community to investigate the properties of RGGs in the case where $d \to \infty$. If the ambitious problem of recognizing if a graph can be realized as a geometric graph is known to be NP-hard Breu and Kirkpatrick [1998], one can take a step back and wonder if a given RGG still carries some spatial information as $d \to \infty$ or if geometry is lost in high-dimensions, a problem known as geometry detection. More precisely, this problem consists in testing if a given graph has been sampled from a Erdös-Renyi random graph or from the RGG model. This question has gained a lot of interest in the past years and we can mention in particular the important contributions from Brennan et al. [2020], Bubeck et al. [2016], Liu et al. [2021]. The proof presented in these papers make use of advanced results from probability, statistics, optimal transport, combinatorics or information theory, placing RGGs at the intersection of a large span of research communities. We can mention in particular that in the dense regime, the phase transition for geometry detection occurs at the regime at which Wishart matrices becomes indistinguishable from GOEs (Gaussian Orthogonal Ensemble) in total variation distance. This question has been investigated in more general settings such as in Bubeck and Ganguly [2015] or Bourguin et al. [2021]. During this thesis, I wrote a survey paper (that will not be discussed further in this manuscript) on the interesting mathematical questions related to RGGs and their extensions.

> Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. The Random Geometric Graph: Recent developments and perspectives. 2022. URL `https://hal.archives-ouvertes.fr/hal-03622277`

Non-parametric estimation in RGGs. In another line of work, De Castro et al. [2019] tackles a non-parametric estimation task in RGGs. Their work contributes to the broader issue of estimation in graphon models. In Tang et al. [2013], the authors prove that spectral methods can recover the matrix formed by graphon evaluated at latent points up to an orthogonal transformation, assuming that graphon is a positive definite kernel (PSD). Going further, algorithms have been designed to estimate graphons, as in Klopp et al. [2017] which provide sharp rates for the SBM. Contrary to the previous works, the paper De Castro et al. [2019] provides a non-parametric algorithm to estimate RGGs on the Euclidean sphere $\mathbb{S}^{d-1}$, without PSD assumption. In their model, they consider n points $X_1, X_2, \dots, X_n$ sampled uniformly and independently on $\mathbb{S}^{d-1}$ and an edge is set between nodes i and j (where $i, j \in [n]$, $i \neq j$) with independent probability $\mathbf{p}(\langle X_i, X_j \rangle)$, where the unknown function $\mathbf{p} : [-1, 1] \to [0, 1]$ is called the *envelope function*. This RGG is a graphon model with a symmetric kernel W given by $W(x, y) = \mathbf{p}(\langle x, y \rangle)$. They show that one can associate to the graphon W an integral operator $\mathbb{T}_W$. The operator $\mathbb{T}_W$ is Hilbert-Schmidt and it has a countable number of bounded eigenvalues $\lambda(\mathbb{T}_W)$ with zero as only accumulation point. The eigenfunctions $(\phi_k)_{k \geq 0}$ of $\mathbb{T}_W$ have the remarkable property that they do not depend on $\mathbf{p}$ (cf. Dai and Xu [2013] Lemma 1.2.3): they are given by the real Spherical Harmonics. One can show that the following spectral decomposition holds

$$\mathbf{p}(t) = \sum_{l \geq 0} p_l^* \phi_k(t) \,,$$

where $\lambda(\mathbb{T}_W) = \{p_0^*, p_1^*, \dots, p_1^*, \dots, p_l^*, \dots, p_l^*, \dots\}$ meaning that each eigenvalue p_l^* has a known multiplicity d_l. This decomposition shows that a plug-in estimation of the envelope function $\mathbf{p}$ can be used if we are able to estimate the eigenvalues of the operator $\mathbb{T}_W$. In De Castro et al. [2019], the authors prove that under some regularity condition on the envelope function, the spectrum of the adjacency matrix of the graph (correctly normalized) converges towards $\lambda(\mathbb{T}_W)$ with respect to the δ_2 metric defined for any square-summable sequences of reals x and y by

$$\delta_2^2(x, y) := \inf_{\sigma \in \mathfrak{S}} \sum_i (x_i - y_{\sigma(i)})^2 \,,$$

where $\mathfrak{S}$ is the set of permutations of natural numbers.
This approach has two major drawbacks: (i) First the algorithm proposed to estimate the envelope function proposed has a complexity that grows exponentially with the chosen resolution level R. (ii)

Secondly, they work with the restrictive condition of independent sampling of the latent positions.

The Markov Random Geometric Graph. In our work

> Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. Markov random geometric graph, MRGG: A growth model for temporal dynamic networks. *Electron. J. Stat.*, 16 (1):671–699, 2022. doi: 10.1214/21-ejs1969. URL https://doi.org/10.1214/21-ejs1969

we make a first step to address both issues by introducing the Markov RGG (MRGG): a growth model for RGG on $\mathbb{S}^{d-1}$ where latent positions are sampled using a Markovian dynamic. More precisely, the distribution of the new latent point X_i is given by $P(X_{i-1}, \cdot)$ where P is a Markov kernel that needs to be estimated. We consider an isotropic Markov sampling scheme meaning that from a current latent position X_i, the next latent point is defined by

$$X_i = r_i \times X_{i-1} + \sqrt{1 - r_i^2} \times Y_i,$$

where

- $Y_i \in \mathbb{S}^{d-1}$ is a unit vector sampled uniformly, orthogonal to X_{i-1},
- $r_i \in [-1, 1]$ encodes the distance between X_{i-1} and X_i. r_i is sampled from a distribution $f_{\mathcal{L}} : [-1, 1] \to [0, 1]$, called the *latitude function*.

Working with this model, our contributions presented in Chapter 3 are the following. (i) We present a polynomial time algorithm based on an *ad hoc* Hierarchical Clustering Algorithm to estimate the envelope function $\mathbf{p}$ and we provide theoretical guarantees for the correctness of our approach when the optimal resolution level is known. (ii) We propose a model selection procedure based on the slope heuristic to estimate a resolution level achieving a relevant bias/variance tradeoff. (iii) We prove that we can estimate consistently the Gram matrix of the latent positions $G^* = n^{-1}\big(\langle X_i, X_j\rangle\big)_{i,j\in[n]}$ in Frobenius norm. (iv) The latter theoretical result motivates the estimation of the latitude function $f_{\mathcal{L}}$ using a kernel density estimator from the approximations obtained of the consecutive latent distances $\big(r_i\big)_{i\in\{2,\dots,n\}} = \big(\langle X_{i-1}, X_i\rangle\big)_{i\in\{2,\dots,n\}}$. (v) We prove that the knowledge of the latent distances are enough to solve link prediction tasks. Hence, based on the above mentioned estimate of the envelope function $\mathbf{p}$, the latitude function $f_{\mathcal{L}}$ and the latent distances G^*, we propose a method to estimate the probability of connection between nodes already present in the graph and the newcomer. (vi) We provide a testing procedure to determine if the given graph hides some latent non-trivial Markovian dynamic or if the nodes have been sampled independently and uniformly on $\mathbb{S}^{d-1}$. (vii) Last but not least, we provide extensive numerical simulations supporting the correctness of our methods.

The theoretical analysis of algorithms used to achieve estimation tasks in dynamic random graph models require powerful probabilistic tools. The MRGG model is no exception to this rule and our proof scheme in Chapter 3 put us in front of the necessity to use a concentration inequality for a U-statistic in a dependent framework. The few existing articles in the literature that addressed this difficult problem considered assumptions that did not fit our framework. This led us to establish a new concentration result for U-statistics for Markov chains.

1.2 Probabilistic tools with dependent random variables and applications

1.2.1 Concentration inequality for U-statistics in a dependent framework (Chapter 4)

In this section, we present a new result for the concentration of U-statistics in a dependent framework. These contributions are fully described in Chapter 4 or in the following paper.

Duchemin et al. [2022b] Quentin Duchemin, Yohann De Castro, and Claire Lacour. Concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. *Bernoulli*, 2022b. URL https://hal.archives-ouvertes.fr/hal-03014763

Context and previous works. Concentration of measure has become ubiquitous in the Statistics and Machine Learning communities. Additionally to random graphs (cf. Chapters 2 and 3), we can mention applications of concentration for model selection (see Massart [2007] and Lerasle et al. [2016]), statistical learning (see Clémençon et al. [2020]) or online learning (see Wang et al. [2012]). Important contributions in this field are those concerning U-statistics. A U-statistic of order m is a sum of the form

$$\sum_{1\leq i_1<\cdots<i_m\leq n} h_{i_1,\dots,i_m}(X_{i_1},\dots,X_{i_m}),$$

where $X_1,\dots,X_n$ are random variables taking values in a measurable space (E,Σ) and where $h_{i_1,\dots,i_m}$ are measurable functions of m variables $h_{i_1,\dots,i_m}:E^m\to\mathbb{R}$.
One important exponential inequality for U-statistics was provided by Arcones and Giné [1993] using a Rademacher chaos approach. Their result holds for bounded and canonical (or degenerate) kernels, namely satisfying for all $i_1,\dots,i_m\in[n]:=\{1,\dots,n\}$ with $i_1<\cdots<i_m$ and for all $x_1,\dots,x_m\in E$,

$$\left\|h_{i_1,\dots,i_m}\right\|_\infty<\infty \quad\text{and}\quad \forall j\in[1,n]\,,\ \mathbb{E}_{X_j}\left[h_{i_1,\dots,i_m}(x_1,\dots,x_{j-1},X_j,x_{j+1},\dots,x_m)\right]=0\,.$$

They proved that in the degenerate case, the convergence rates for U statistics are expected to be $n^{m/2}$. Relying on precise moment inequalities of Rosenthal type, Giné et al. [2000] improved the result from Arcones and Giné [1993] by providing the optimal four regimes of the tail, namely Gaussian, exponential, Weibull of orders $2/3$ and $1/2$. When the kernels are unbounded, it was shown that some results can be extended provided that the random variables $h_{i_1,\dots,i_m}(X_{i_1},\dots,X_{i_m})$ have sufficiently light tails [see for example Eichelsbacher and Schmock, 2003, Theorem 3.26]
All the above mentioned results consider that the random variables $(X_i)_{i\geq1}$ are independent. The asymptotic behaviour of U-statistics in a dependent setup has already been investigated by several papers [see for example Bertail and Clémençon, 2011, Eichelsbacher and Schmock, 2001]. The main works providing concentration inequality for U-statistics with dependent random variables are Borisov and Volodko [2015], Han [2018] and Shen et al. [2020]. All these papers consider a fixed kernel (namely $h\equiv h_{i_1,\dots,i_m}$ for all $i_1,\dots,i_m$) defined on $\mathbb{R}^d$ with strong regularity conditions. For the first time, we consider in this thesis time dependent kernel functions which makes the theoretical analysis more challenging since the standard splitting method can be unworkable (cf. Section 4.2.5).

Assumptions.
We consider a Markov chain $(X_i)_{i\geq1}$ with a transition kernel $P:E\times E\to\mathbb{R}$ taking values in a measurable space (E,Σ), and we introduce measurable functions $h_{i,j}:E^2\to\mathbb{R}$. Our goal is to study the concentration properties of the U-statistic

$$U_{\text{stat}}(n)=\sum_{1\leq i<j\leq n}\left(h_{i,j}(X_i,X_j)-\mathbb{E}\left[h_{i,j}(X_i,X_j)\right]\right).$$

We work with the following set of assumptions.

1. **Uniform ergodicity**: *The Markov chain $(X_i)_{i\geq1}$ is assumed to be uniformly ergodic with stationary measure π.*
2. **Bounded transition kernel**: *There exist some $\delta_M>0$ and a probability measure ν such that*
$$\forall x\in E,\ \forall A\in\Sigma,\quad P(x,A)\leq\delta_M\nu(A).$$
3. **π-canonical and bounded kernels**: *For all $i,j\in[n]$, $h_{i,j}:E\times E\to\mathbb{R}$ is mesurable, bounded and π-canonical, namely*
$$\forall x,y\in E,\quad \mathbb{E}_\pi[h_{i,j}(X,x)]=\mathbb{E}_\pi[h_{i,j}(X,y)]=\mathbb{E}_\pi[h_{i,j}(x,X)]=\mathbb{E}_\pi[h_{i,j}(y,X)].$$

This commom expectation is denoted $\mathbb{E}_\pi[h_{i,j}]$.

4. **Technical assumption**: *At least one of the following conditions holds*

   (i) For all $i, j \in [n]$, $h_{i,j} \equiv h_{1,j}$, *i.e. the kernel* $h_{i,j}$ *does not depend on* i.

   (ii) The initial distribution of the chain is absolutely continuous with respect to π *and its density has a finite* p*-th moment for some* $p \in (1, \infty]$.

In Chapter 4, we provide several important examples of Markov chains satisfying our set of assumptions.

Main results. For the first time, we provide in this thesis a concentration inequality for U-statistics of order two in a dependent framework with kernels that may depend on the indexes of the sum and that are not assumed to be symmetric or smooth.
First, we prove a Hoeffding-type concentration result which holds without any condition (or under a mild condition) on the initial distribution of the chain. Assuming that the Markov chain $(X_i)_{i\geq 1}$ is stationary (which means that X_1 is distributed according to π), we prove a Bernstein-type concentration inequality which leads to a better convergence speed. Our main results are summarized in Theorem 2. Our concentration inequality involves quantities B_n and C_n that can be interpreted as standard deviation terms and we refer to Chapter 4 for their precise definitions. In order to read directly the dominant terms in our concentration inequality from Theorem 2, let us highlight that one can always bound coarsely B_n and C_n as follows

$$B_n \leq \sqrt{n}A \quad \text{and} \quad C_n \leq nA \quad \text{where} \quad A := 2 \max_{i,j} \|h_{i,j}\|_\infty.$$

Theorem 2

We consider that the Assumptions 1 to 4 are satisfied. Then there exist two constants $\beta, \kappa > 0$ *such that for any* $u > 0$, *it holds with probability at least* $1 - \beta e^{-u} \log n$,

$$U_{\text{stat}}(n) \leq \kappa \log(n) \Big(\left[C_n + A \log(n) \sqrt{n} \right] \sqrt{u} + \left[A + B_n \sqrt{n} \right] u + \left[2A\sqrt{n} \right] u^{3/2} + A \left[u^2 + \alpha_n \right] \Big),$$

where $\alpha_n = \begin{cases} \log(n) & \text{if the chain } (X_i)_{i\geq 1} \text{ is stationary} \\ n & \text{otherwise} \end{cases}$.

If Assumption 4.(i) *holds, one can remove* C_n *in the previous inequality.*

In Chapter 4, we motivate the use of index-dependent kernels presenting two specific examples borrowed from the fields of information retrieval and of homogeneity tests. Considering a particular case, we show that our Bernstein inequality (obtained when $\alpha_n = \log(n)$ in Theorem 2) can lead to significantly smaller convergence rates.

In the three following sections, we describe three important applications to Statistics and Machine Learning of Theorem 2. These contributions are presented in Chapter 5 and in the following paper.

Duchemin et al. [2022a] Quentin Duchemin, Yohann De Castro, and Claire Lacour. Three rates of convergence or separation via U-statistics in a dependent framework. *JMLR*, 2022a. URL https://hal.archives-ouvertes.fr/hal-03603516

1.2.2 Estimation of spectra of signed integral operator with MCMC algorithm (Chapter 5 Sec.5.3)

Context. In learning theory such as in Principal Component Analysis (PCA) or some manifold methods [cf. Rosasco et al., 2010], estimating the eigenvalues and/or the eigenvectors of data-dependent matrices is essential. It appears that these matrices can often be interpreted as the empirical versions of continuous objects such as integral operators. As highlighted in Rosasco et al. [2010], the theoretical analysis of the above mentioned learning algorithms requires to quantify the difference between the

eigen-structure of the empirical operators and their continuous counterparts. In this thesis, we study the convergence of sequence of spectra of kernel matrices towards the spectrum of some integral operator. Previous important works may include Adamczak and Bednorz [2015a] and, as far as we know, they all assume that the kernel is of positive-type (*i.e.,* giving an integral operator with non-negative eigenvalues). For the first time, we prove a non-asymptotic result of convergence of spectra for kernels that are not of positive-type. We further prove that *independent Hastings algorithms* are valid sampling schemes to apply our result.

Result. We consider a Markov chain $(X_n)_{n\geq 1}$ on E satisfying Assumptions 1 and 2 with stationary distribution π, and some kernel $h : E \times E \to \mathbb{R}$ satisfying the following assumptions.

> *$h : E\times E \to \mathbb{R}$ is a bounded and symmetric function square integrable with respect to $\pi\otimes\pi$. Moreover there exist continuous functions $\phi_r : E \to \mathbb{R}$, $r \in I$ (where $I = \mathbb{N}$ or $I = 1, \dots, N$) that form an orthonormal basis of $L^2(\pi)$ and a sequence of real numbers $(\lambda_r)_{r\in I} \in \ell_2$ such that it holds pointwise*
>
> $$h(x, y) = \sum_{r\in I} \lambda_r \phi_r(x)\phi_r(y),$$
>
> *with $\sup_{r\in I} \|\phi_r\|_\infty^2 < \infty$ and $\sup_{x\in E} \sum_{r\in I} |\lambda_r|\phi_r(x)^2 < \infty$.*

We can associate to h the kernel of a linear operator $\mathbf{H}$ defined by

$$\mathbf{H}f(x) := \int_E h(x, y)f(y)d\pi(y).$$

This is a Hilbert-Schmidt operator on $L^2(\pi)$ and thus it has a real spectrum consisting of a square summable sequence of eigenvalues and we denote the eigenvalues of $\mathbf{H}$ by $\lambda(\mathbf{H}) := (\lambda_1, \lambda_2, \dots)$. For some $n \in \mathbb{N}^*$, we consider $\mathbf{H}_n := \frac{1}{n} (h(X_i, X_j))_{1\leq i,j\leq n}$ with eigenvalues $\lambda(\mathbf{H}_n)$.
In Section 5.3, we prove that the spectrum of $\mathbf{H}_n$ converge towards the spectrum of the integral operator $\mathbf{H}$ as $n \to \infty$. More precisely, there exist constants C, D such that for n large enough it holds with probability at least $1 - D/\sqrt{n}$,

$$\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 \leq \frac{C \log n}{\sqrt{n}} + 8 \sum_{i>\lceil n^{1/4}\rceil, i\in I} \lambda_i^2. \tag{1.1}$$

Let us point out that the proof scheme of this result generalizes an approach already exploited in Chapter 3.

Application. We are now given some kernel h and a probability measure π satisfying the previous assumptions. We aim at computing the eigenvalues of the integral operator $\mathbf{H}$ associated to h. π often does not admit a closed-form expression, a situation that typically arises in a Bayesian context where π is some posterior distribution. A standard way to cope with this issue is to rely on MCMC methods. In Section 5.3, we adopt this approach and we consider the specific case where E is a bounded subset of $\mathbb{R}^k$ equipped with the Borel σ-algebra $\mathcal{B}(E)$. We consider a probability density q on E, called the proposal distribution. We assume that the measure π on E admits a density f_π with respect to the Lebesgue measure λ_{Leb} on E and that it holds

$$\forall y \in E, \quad f_\pi(y), q(y) > 0 \quad \text{and} \quad \frac{q(y)}{f_\pi(y)} > \beta \ \text{ for some } \beta > 0\,.$$

In this context, we prove in Section 5.3 that a Markov chain $(X_i)_{i\geq 1}$ obtained from an independent Hastings algorithm with proposal distribution $q\lambda_{Leb}$ satisfies Assumptions 1 and 2. We deduce that one can estimate the eigenvalues of $\mathbf{H}$ by computing the ones of $\mathbf{H}_n$ and Eq.(1.1) quantifies the distance between both spectra.

1.2.3 Generalization bounds for online learning with pairwise loss function (Chapter 5 Sec.5.4)

Context. In Machine Learning, batch algorithms accumulate data over a period of time and only train the models once the data acquisition process is completed. Batch learning has some limitations especially when we receive data as a continuous flow (e.g., stock prices) and we need to adapt to changes rapidly, or for large scale learning problems where their computational cost can be prohibitive. Online algorithms have been designed to efficiently solve learning problems in such situations: they deal with data coming on fly and try to improve the learned model along time based on the new observations. One way to analyze the performance of online learning algorithms is to consider the notion of *regret* which compares the difference between the payoff obtained by the learning algorithm and the payoff that would have been obtained by taking the optimal decision at each time step. In the last decade, researchers were not only interested in the notion of regret but looked at online learning algorithms through a different lens by wondering how they could generalize on future data. This question only makes sense if we assume that the sequence of examples comes from some stochastic process. As asked in Agarwal and Duchi [2012], *"if the sequence of examples are generated by a stochastic process, can the online learning algorithm output a good predictor for future samples from the same process?"*
The generalization performance of online learning algorithms with univariate loss functions has been so far well studied with both i.i.d. or dependent observations. Working with pairwise loss functions, generalization bounds have been obtained with i.i.d. data but this thesis provides one of the first result for the case of dependent observations. Figure 1.2 depicts the framework that we consider in Section 5.4.

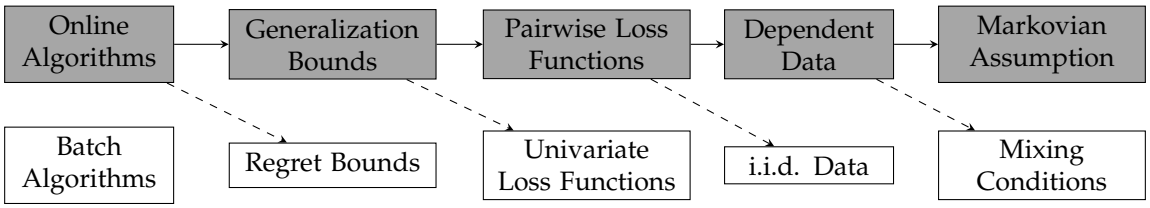


Figure 1.2: Positioning our contributions in the existing literature for the analysis of online algorithms.

Results. Inspired by the ranking problem, we consider a function $f : E \to \mathbb{R}$ which defines the ordering of the objects in E. We aim at finding a relevant approximation of the ordering of the objects in E by selecting a function h (called a *hypothesis* function) in a space $\mathcal{H}$ based on the observation of the random sequence $(X_i, f(X_i))_{1\leq i\leq n}$ where $(X_i)_{i\geq 1}$ is a reversible Markov chain satisfying Assumptions 1 and 2. To measure the performance of a given hypothesis $h : E \times E \to \mathbb{R}$, we use a pairwise loss function of the form $\ell(h, X, U)$. Typically, one could use the *misranking loss* defined by

$$\ell(h, x, u) = \mathbb{1}_{\{(f(x)-f(u))h(x,u)<0\}},$$

which is 1 if the examples are ranked in the wrong order and 0 otherwise. The goal of the learning problem is to find a hypothesis h which minimizes the *expected misranking risk*

$$\mathcal{R}(h) := \mathbb{E}_{(X,X')\sim\pi\otimes\pi}\Big[\ell(h, X, X')\Big].$$

In the context of online learning, at each time step t the algorithm chooses some hypothesis $h_t \in \mathcal{H}$ based on the sequence of observations $(X_i, f(X_i))_{i\leq t}$ up to time t. We work with the following assumptions.

- $(\mathcal{H}, \|\cdot\|_\infty)$ *is compact and satisfies*

$$\log \mathcal{N}(\mathcal{H}, \eta) = \mathcal{O}(\eta^{-\theta}),$$

where $\mathcal{N}(\mathcal{H}, \eta)$ is the L^∞-covering number of $\mathcal{H}$ and where $\theta > 0$.

- *The loss function $\ell : \mathcal{H} \times E \to [0, 1]$ is such that*

$$\ell(h, x_1, x_2) = \phi(f(x_1) - f(x_2), h(x_1, x_2)),$$

where $\phi : \mathbb{R}\times \to [0, 1]$ is Lipschitz with respect to its second coordinate.

Our contributions presented in Section 5.4 are the following:

1. We introduce a new *average paired empirical risk*, denoted by $\mathcal{M}^n := \mathcal{M}^n(h_1, \dots, h_{n-1-b_n})$, that can be computed in practice. It depends on the key quantity b_n that can be interpreted as a forgetting factor. b_n scales with $\log n$ and its definition involves a constant accounting for the mixing properties of the chain.

2. We give non-asymptotic error bounds between $\mathcal{M}^n$ and the true average risk, namely denoting $c_n = \lfloor cn \rfloor$ for some $c \in (0,1)$,

$$\left| \frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \mathcal{R}(h_{t-b_n}) - \mathcal{M}^n \right| = \mathcal{O}_{\mathbb{P}} \left[\frac{\log(n) \log(\log n)}{n^{\frac{1}{2+\theta}}} \right].$$

3. We convert a regret bound of an online learner into a control of the excess risk. More precisely, considering an online learner that achieves a regret bound $\mathfrak{R}_n$ i.e. such that

$$\mathcal{M}^n \leq \inf_{h \in \mathcal{H}} \left\{ \mathcal{M}^n(h, \dots, h) \right\} + \mathfrak{R}_n,$$

   we show that the average risk of the ensemble of hypotheses $(h_t)_{t \geq 1}$ satisfies

$$\frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \mathcal{R}(h_{t-b_n}) - \min_{h \in \mathcal{H}} \mathcal{R}(h) = \mathcal{O}_{\mathbb{P}} \left[\frac{\log(n) \log(\log n)}{n^{\frac{1}{2+\theta}}} + \mathfrak{R}_n \right].$$

4. We build a hypothesis selection procedure that outputs some $\widehat{h} \in \{h_t,\ t \in [n]\}$ achieving this average risk.

1.2.4 Adaptive goodness-of-fit tests in a density model (Chapter 5 Sec.5.5)

Context. Several works have already proposed goodness-of-fit tests for the density of the stationary distribution of a sequence of dependent random variables and we can mention for example Bai [2003], Chwialkowski et al. [2016], Li and Tkacz [2001]. In all the above mentioned papers, asymptotic properties of the test statistic are derived but no non-asymptotic analysis of the methods is conducted. As far as we know, we provide for the first time a non-asymptotic condition on the classes of alternatives ensuring that the statistical test reaches a prescribed power working in a dependent framework.

Result. We consider a Markov chain $X_1, \dots, X_n$ with stationary distribution π with density f with respect to the Lebesgue measure on $\mathbb{R}$ satisfying Assumptions 1 and 2. Let f_0 be some given density in $L^2(\lambda_{Leb})$ and let α be in $]0,1[$. Assuming that f belongs to $L^2(\lambda_{Leb})$, we construct a level α test of the null hypothesis "$f = f_0$" against the alternative "$f \neq f_0$" from the observation $(X_1, \dots, X_n)$. The test is based on the estimation of $\|f - f_0\|_2^2$ that is $\|f\|_2^2 + \|f_0\|_2^2 - 2\langle f, f_0 \rangle$. $\langle f, f_0 \rangle$ is usually estimated by the empirical estimator $\sum_{i=1}^n f_0(X_i)/n$ and the cornerstone of our approach is to find a way to estimate $\|f\|_2^2$. We follow the work of Fromont and Laurent [2006] and we introduce a set $\{S_m, m \in \mathcal{M}\}$ of linear subspaces of $L^2(\lambda_{Leb})$. In Section 5.5, we consider three different collections of linear subspaces $\{S_m, m \in \mathcal{M}\}$, namely constant piecewise functions, scaling functions and trigonometric polynomials. For all m in $\mathcal{M}$, let $\{p_l, l \in \mathcal{L}_m\}$ be some orthonormal basis of S_m. The variable

$$\widehat{\theta}_m = \frac{1}{n(n-1)} \sum_{l \in \mathcal{L}_m} \sum_{i \neq j = 1}^{n} p_l(X_i) p_l(X_j)$$

estimates $\|\Pi_{S_m}(f)\|_2^2$ where Π_{S_m} denotes the orthogonal projection onto S_m. Then $\|f - f_0\|_2^2$ can be approximated by

$$\widehat{T}_m = \widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n} \sum_{i=1}^{n} f_0(X_i),$$

for any m in $\mathcal{M}$. Denoting by $t_m(u)$ the $(1-u)$ quantile of the law of $\widehat{T}_m$ under the hypothesis "$f = f_0$" and considering

$$u_\alpha = \sup_{u\in]0,1[} \mathbb{P}_{f_0}\left(\sup_{m\in\mathcal{M}} (\widehat{T}_m - t_m(u)) > 0\right) \leq \alpha,$$

we introduce the test statistic T_α defined by

$$T_\alpha = \sup_{m\in\mathcal{M}} (\widehat{T}_m - t_m(u_\alpha)).$$

The test consists in rejecting the null hypothesis if T_α is positive. This approach can be read as a multiple testing procedure. Indeed, for each m in $\mathcal{M}$, we construct a level u_α test of the null hypothesis "$f = f_0$" by rejecting this hypothesis if $\widehat{T}_m$ is larger than its $(1-u_\alpha)$ quantile under the hypothesis "$f = f_0$". We thus obtain a collection of tests and we decide to reject the null hypothesis if for some of the tests of the collection this hypothesis is rejected.
In Section 5.5, we provide an upper-bound on the so-called *separation rate* for specific classes of alternatives that include some Besov bodies. Let us recall that the separation rate associated with a class of alternatives $\mathcal{B}$ and $\gamma \in (0,1)$ is defined as the smallest real $\rho > 0$ such that for any $f_1 \in \mathcal{B}$ with $\|f_0 - f_1\|_2 > \rho$, the power of our test for the alternative "$f = f_1$" is larger than $1-\gamma$. In Section 5.5, we prove that the separation rate is upper-bounded by

$$\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}},$$

where the parameter $s > 0$ quantifies the regularity of the densities belonging to our class of alternatives. Note that in the i.i.d. setting, the adaptive minimax rate of testing is known to be of order $(\sqrt{\log\log n}/n)^{2s/(4s+1)}$ (see Ingster [1993]). At the end of Section 5.5, we provide numerical experiments comparing our approach with the Kolmogorov-Smirnov test and the χ^2-test for several different Markov chains.

We now come back to one of the main initial motivation of this PhD thesis: the problem of link prediction in random graphs. In Chapters 2 and 3, we proposed to tackle this question in graphon models where nodes join the graph along time with a latent representation depending on the last entrant. We will assume in the next section that latent representations of nodes $X_i \in \mathbb{R}^d$ are observed and that the graphon belongs to some high-dimensional parametric class based on logistic regression. Such high-dimensional setting arises in a large number of Machine Learning problems and often makes the estimation task ill-posed. To cope with this issue, the analyst assumes some structure on the problem (typically some sparsity assumption), performs a model selection step and then estimates the parameter of interest using the selected model. Since data was used to select the model, using the standard machinery for inference can lead to undesirable statistical properties. Post-selection inference (PSI) aims at addressing this problem by taking into account the selection event to provide valid inference procedures. In Chapter 6, we push further the current state of knowledge for PSI methods in generalized linear models with a specific focus on the sparse logistic regression.

1.3 Selective inference in the sparse logistic regression (Chapter 6)

Motivation. Still motivated by link prediction problems in random graphs, we consider now a framework where for each node $i \in [n]$ of a simple and undirected graph of size n, we are given side information, a vector of observations $X_i \in \mathbb{R}^d$. Given observations X_i, X_j about nodes i and j of the network, we assume that nodes i and j are connected with probability $W(X_i, X_j)$ for some symmetric map $W : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$. Inspired by the work from Berthet and Baldin [2020], we consider that the graphon W belongs to some high-dimensional class based on logistic regression. Denoting by $\sigma : x \mapsto (1+\exp(-x))^{-1}$ the sigmoid function, we consider that there exists some $\Theta^* \in \mathbb{R}^{d\times d}$ such that $W(X_i, X_j) = \sigma(X_i^\top \Theta^* X_j)$. Our goal is to estimate the unknown matrix Θ^* based on the observation of the adjacency matrix of the graph A, and on the known explanatory variables $\mathbb{X} := [X_1 \ \dots \ X_n] \in \mathbb{R}^{d\times n}$ and to conduct inference procedures (through hypothesis tests or confidence intervals). We have the prior that Θ^* is sparse, meaning that only a small subset of the observed d covariates drive that connection between two nodes in the graph. In this situation, the classical approach adopted by statisticians

consist to follow a three-stage protocol.

- *Model selection*: Based on the observed data, the statistician selects a subset of active entries in Θ. One standard approach is to compute the ℓ_1-penalized maximium likelihood estimator

$$\widehat{\Theta}^{\lambda} \in \arg\min_{\Theta \in \mathbb{R}^{d\times d}} \big\{ -\log \mathbb{P}_{\Theta}(A \,|\, \mathbb{X}) + \lambda \|\Theta\|_1 \big\}, \tag{1.2}$$

  which can be written as a classical logistic regression. Indeed, by writing $\mathrm{vec}(B) \in \mathbb{R}^{p^2}$ the vectorized form of a matrix $B \in \mathbb{R}^{p\times p}$, we have that

$$X_j^{\top} \Theta^* X_i = \mathrm{Tr}(X_i X_j^{\top} \Theta^*) = \langle \mathrm{vec}(X_i X_j^{\top}), \mathrm{vec}(\Theta^*) \rangle.$$

  Hence, using obvious notations, Eq.(1.2) is equivalent to

$$\widehat{\vartheta}^{\lambda} \in \arg\min_{\vartheta \in \mathbb{R}^{d^2}} \big\{ -\log \mathbb{P}_{\vartheta}(Y \,|\, \mathbf{X}) + \lambda \|\vartheta\|_1 \big\},$$

  where $Y = \mathrm{vec}(A), \mathbf{X} = [\mathrm{vec}(X_1 X_1^{\top})\ \mathrm{vec}(X_1 X_2^{\top})\ \ldots\ \mathrm{vec}(X_n X_n^{\top})] \in \mathbb{R}^{d^2 \times n^2}$ and where $\widehat{\vartheta}^{\lambda} = \mathrm{vec}(\widehat{\Theta}^{\lambda})$. Then, we define the set of the active entries as $M := \widehat{M}(Y) := \{i \,|\, \widehat{\vartheta}_i^{\lambda} \neq 0\}$.

- *Estimation*: The statistician computes the Maximum Likelihood Estimator (MLE) using only the variables in M

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^{|M|}} \big\{ -\log \mathbb{P}_{\vartheta(\theta)}(Y \,|\, \mathbf{X}) \big\},$$

  where for any $\theta \in \mathbb{R}^{|M|}$, $\vartheta(\theta) \in \mathbb{R}^{d^2}$ is such that $\vartheta_{-M}(\theta) = 0$ and $\vartheta_M(\theta) = \theta$.

- *Inference*: The statistician conducts hypothesis tests or provides confidence intervals.

Post-selection inference (PSI). The model selection step requires the choice of the hyperparameter λ which is performed in practice by using the data. In this context, applying standard inference methods without taking into account we used the data to select the model will generally leads to undesirable frequency properties (cf. Pötscher [1991]). Post-selection inference aims at solving this problem. The method consists in producing inference procedure considering that the vector of observations Y is distributed according to $\mathbb{P}_{\vartheta^*}(\tilde{Y} \,|\, \mathbf{X}, \{\tilde{Y} \in E_M\})$. In the former conditional distribution, $E_M := \{\tilde{Y} \,|\, M = \widehat{M}(\tilde{Y})\}$ is called the *selection event* and corresponds to the set of all random graphs with vectorized adjacency matrix A given by $\tilde{Y}$ that would have led to the same set of active variables than the graph with adjacency matrix B with $\mathrm{vec}(B) = Y$. PSI in the context of linear regression has known an increasing interest in the result years, in particular thanks to the important breakthrough made by Lee et al. [2016]. In the former paper, the authors prove that in the linear model with gaussian noise, the distribution of the response variable conditional on the selection event is a mixture of truncated multivariate Gaussians. This result is a consequence of the so-called polyhedral lemma and allows to provide exact PSI procedures in this context. Methods for PSI outside of the linear model with gaussianity have been investigated recently and one can mention Fithian et al. [2014], Taylor and Tibshirani [2018], Tian and Taylor [2017], Tian et al. [2018], Tibshirani et al. [2018] just to name a few. Despite the ubiquity of the logistic regression model in applications, it remains one of the setting where existing PSI methods for generalized linear models (GLMs) are either unsuited [cf. Fithian et al., 2014, Section 6.3] or lack theoretical guarantees [cf. Taylor and Tibshirani, 2018].

Contributions. In Chapter 6, (i) we provide a new formulation of the selection event in GLMs shedding light on the essential diffeomorphism Ψ that carries the geometric information of the problem, (ii) we provide a new perspective on post-selection inference in GLMs through the conditional MLE approach of which Ψ is a key ingredient, (iii) we introduce sufficient conditions in GLMs to obtain asymptotically valid PSI procedures based on the conditional MLE approach. Thereafter, we focus on the specific case of logistic regression: (iv) we prove - under some assumptions - that the sufficient conditions from (iii) are satisfied for the logistic model. (v) This allows us to give asymptotically valid PSI procedures for logistic regression and we support are theoretical results with simulations. (vi) Finally, we present an extensive comparison between our work and the heuristic from Taylor and Tibshirani [2018] which is currently considered as the best to use in the context of the sparse logistic regression [cf. Fithian et al., 2014, Section 6.3].

This work on PSI in GLMs is presented in Chapter 6 and corresponds to the following paper.

Duchemin and De Castro [2022] Quentin Duchemin and Yohann De Castro. A new procedure for Selective Inference with the Generalized Linear Lasso. 2022. URL https://hal.archives-ouvertes.fr/hal-03622196

Chapter 2

Reliable Temporal Prediction in the Markov Stochastic Block Model

Chapter Abstract

In this chapter, we introduce the Markov Stochastic Block Model: an extension of SBMs where communities of the nodes are assigned through a Markovian dynamic. We show how MSBMs can be used to detect dependence structure in growing graphs and we provide methods to solve the so-called link prediction and collaborative filtering problems. We make our approaches robust with respect to the outputs of the clustering algorithm and we propose a model selection procedure. Our methods can be applied regardless of the algorithm used to recover communities in the network. In this paper, we use a recent Semi-Definite Programming (SDP) method to infer the hidden communities and we provide theoretical guarantees. In particular, we identify the relevant signal-to-noise ratio (SNR) in our framework and we prove that the misclassification error decays exponentially fast with respect to this SNR.

Chapter Content

2.1 Introduction

2.1.1 Context

Large random graphs have been very popular in the last decade since they are powerful tools to model complex phenomena like interactions on social networks Yang et al. [2011] or the spread of a disease Ahmad and Xu [2017]. In practical cases, detecting communities of well connected nodes in a graph is a major issue, motivating the study of the Stochastic Block Model (SBM). In this model, each node belongs to a particular community and edges are sampled independently according to a probability depending of the communities of the nodes. Aiming at progressively bridging the gap between models and reality, time evolving SBMs have been recently introduced. In Matias and Miele [2015], a Stochastic Block Temporal Model is considered where the temporal evolution is modeled through a discrete hidden Markov chain on the nodes membership and where the connection probabilities also evolve through time. In Pensky and Zhang [2017], connection probabilities between nodes are functions of time, considering a maximum number of nodes that can switch their communities between two consecutive time steps. Following the work of Karrer and Newman [2011], Lei and Rinaldo [2015] study the Degree Corrected Stochastic Block Model where the degree of the nodes can vary within the same community. They show that for the relatively sparse case (i.e. when the maximum expected node degree is of order $\log(n)$ or higher), the proportion of misclassified nodes tends to 0 with a probability that goes to 1 when the number of nodes n increases using spectral clustering. This result inspired the recent paper Keriven and Vaiter [2022] which considers a Dynamic Stochastic Block Model where the communities can change with time. They provide direct connection between the density of the graph and its smoothness (which measures how much the graph changes with time). Several other dynamic variants of the SBM have been proposed so far like in Xu [2014] where the presence of an edge at the time step $t+1$ directly depends on its presence or absence at time t. The above mentioned works are mainly considering SBMs where membership of nodes or edges can evolve with time, but only few papers are interested in growth model for SBMs (meaning that the size of the graph increases along time) and we aim at filling this gap.

2.1.2 Standard SBM and tools for community detection

Different recovery requirements have been studied in the SBM. Exact recovery defines the ability to recover the true partition of the nodes as the size of graph tends to $+\infty$ while weak recovery aims at asymptotically recovering correctly a fixed and a non trivial fraction of the partition of the nodes. The survey Abbe [2017] gathers the state of the art methods to solve the community detection problem in the SBM which includes in particular belief propagation algorithms Abbe and Sandon [2016] or spectral methods Chin et al. [2015]. Neural networks Shchur and Günnemann [2019], Bayesian approaches Yang et al. [2011] or Maximum Likelihood estimation Celisse et al. [2012] have also been proposed to address the community detection problem. Another powerful and popular tool is Semi-Definite Programming (SDP) which is known to have interesting robustness features Perry and Wein [2017], Fei and Chen [2018]. Recently, Giraud and Verzelen [2019] proposed a SDP method to address community detection by solving a relaxed version of K-means. They get partial recovery bound with a misclassification error that decays exponentially fast with the signal-to-noise ratio. In our paper, we use their method to recover communities in MSBMs. Therefore, we present succinctly their approach in Section 2.8.

2.1.3 Time prediction in SBMs

Time prediction in random graphs. Networks are nice structures to achieve prediction tasks. One of them is link prediction which consists in finding missing links or in inferring the interactions between nodes in the future. Such questions have gained a lot of interest in the recent years leading to both practical Armengol et al. [2015] and theoretical Berthet and Baldin [2020] works. Link prediction in community-based random graph models have already been studied such as in Biswas and Biswas [2016], but they do not consider temporal evolution in their model. Taking into account temporal aspects for link prediction as in Dunlavy et al. [2011], Bu et al. [2019] is a very active research direction of great interest for applications. Another line of work is interested in reaching a better understanding of the reliability of the link prediction methods typically when we work with noisy observations Guimerà and Sales-Pardo [2009], Feng et al. [2012]. In our work, we follow both directions proposing reliable methods to solve link prediction tasks for networks with an underlying temporal dynamic.

Motivations. While previous works mainly consider a fixed number of nodes with an evolving graph where communities or connection probabilities can evolve, the MSBM is a growth model where a new node enters the graph at each time step and its community is drawn from a distribution depending only on the community of its predecessor. Our model could find interesting applications as in the study of bird migrations (see Section 2.6.3) where animals have regular seasonal movement between breeding and wintering grounds. Another possible application of our model is for recommendation systems. Suppose that we have access to the online purchases of some customers. We know for each of them the date and the product ID of each of their purchases. In Section 2.9.3, we explain how our model can be used to address the following tasks. $i)$ cluster the product IDs by category $ii)$ learn the purchasing behavior of each customer $iii)$ use this information to suggest relevant new products to each customer.

2.1.4 Contributions and Outline

Contributions. We show that the MSBM gives a convenient framework to extract reliable time information from the graph using any reasonable clustering algorithm. We propose a hypothesis test to distinguish between classical SBMs and MSBMs and we address link prediction and collaborative filtering problems. We show that the standard plug-in method is highly sensitive to clustering errors. This is the reason why we propose a reliable approach that takes into account potential errors in the estimated communities. To do so, we learn – using the Baum Welch algorithm – the probability that the clustering algorithm predicts community l for a node belonging to community k for any $k, l \in [K]$. Based on these quantities, we also propose a model selection procedure. In our simulations, we use the method from Giraud and Verzelen [2019] to recover communities and as far as we know, we are the first to provide an implementation of their algorithm. From a theoretical point of view, we show that the misclassification error decays exponentially fast with respect to the signal-to-noise ratio (SNR) and we provide regimes where we can estimate consistently the parameters of our model.

Outline. In Section 2.2, we formally define SBMs and we introduce MSBMs. Furthermore, we establish a partial recovery bound and we show that we can consistently estimate the parameters of our model. Then come our three main contributions in Sections 2.3, 2.4 and 2.5 where we show how the MSBM can be used to study growing networks: we present in particular procedures to solve link prediction tasks or collaborative filtering problems. We give heuristics to be robust to potential local clustering errors of the algorithm. Section 2.6 is dedicated to numerical experiments where we propose a method to infer the unknown number of clusters and where we apply our methods on real data.
In the last three sections of this chapter, we provide additional material and proofs.

2.2 Model and Estimation procedures

2.2.1 Presentation of the MSBM

An undirected graph G is defined by a set of nodes V and a set of edges $E \subset V \times V$. For an undirected graph with n nodes, we define the adjacency matrix of this graph $X \in \{0,1\}^{n\times n}$ such that for all $i, j \in [n]$,

$$X_{i,j} = \begin{cases} 1 & \text{if } \{i,j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

Stochastic Block Model. Let us consider $K \geq 2$ communities and a set of n nodes $V = [n]$. The communities $(c_i)_{i\in[n]} \in K^n$ are assigned independently to each node according to a probability distribution $\nu \in [0,1]^K$, $\sum_{k\in[K]} \nu_k = 1$. Stated otherwise, the community c_i of node $i \in [n]$ is randomly sampled from the distribution ν. Considering the symmetric connectivity matrix $Q \in [0,1]^{K\times K}$, the adjacency matrix of the graph $X \in \{0,1\}^{n\times n}$ related to the assignment of the communities $(c_i)_{i\in[n]}$ is defined by

$$X_{i,j} \sim \mathrm{Ber}(Q_{c_i,c_j}),$$

where $\mathrm{Ber(p)}$ indicates a Bernoulli random variable with parameter $p \in [0,1]$. In the standard SBM, X is observed while the latent variables $(c_i)_{i\in[n]}$ are unknown.

For a parameter $\alpha_n \in (0,1)$ varying with the number of nodes n, we will be focused on connectivity matrix of the form

$$Q := \alpha_n Q_0,$$

where $Q_0 \in [0,1]^{K\times K}$ is a matrix independent of n. As highlighted for example in Abbe and Sandon [2015a], the rate of α_n as $n \to \infty$ is a key property to study random graphs sampled from SBMs. Typical regimes are $\alpha_n \sim 1$ (dense regime), $\alpha_n \sim \frac{\log(n)}{n}$ (relatively sparse regime) and $\alpha_n \sim \frac{1}{n}$ (sparse regime).

Markovian assignment of communities in the SBM. We introduce in this paper the Markov Stochastic Block Model (MSBM) which assigns a community to each node using a Markovian dynamic. We start by ordering the n nodes in V and without loss of generality, we consider the increasing order of the integers $1, 2, \dots, n$. For all $i \in [n]$, we denote $C_i \in [K]$ the random variable representing the community of the node i and we consider that they satistify the following assumption.

Assumption A1. $(C_i)_{i\in[n]}$ is a positive recurrent Markov chain on the finite space $[K]$ with stationary measure π, with transition matrix $P \in \mathbb{R}^{K\times K}$ and initial distribution π. K is independent of n.

We assign communities as follows:

$C_1 \sim \pi$
For $i = 1 \ \dots \ (n-1)$ Do
  $C_{i+1} \sim P_{C_i,:}$
EndFor.

Once the community of each node is assigned, we draw an edge between the nodes i and j with probability Q_{C_i,C_j}

$$X_{i,j} \sim \text{Ber}(Q_{C_i,C_j}) \quad \text{with} \quad Q := \alpha_n Q_0.$$

Here, $Q_0 \in [0,1]^{K\times K}$ is independent of n and $\alpha_n \in (0,1)$ is varying with n. Figure 2.1 presents a graphical representation of our model. We observe the adjacency matrix X but the latent variables $(C_i)_{i\in[n]}$ are unknown.

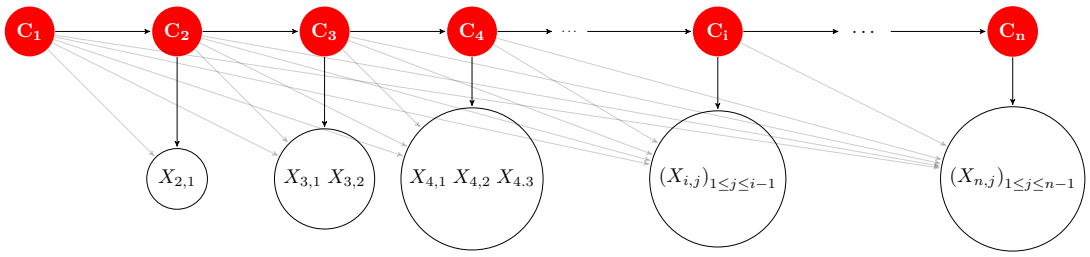


Figure 2.1: Graphical model presenting the SBM with Markovian assignment of the communities.

The following quantities will be crucial in the definition of the signal-to-noise ratio (SNR)

$$L := \|Q_0\|_\infty, \quad \pi_m := \min_{c\in[K]} \pi(c), \quad D^2 := \min_{l\neq k} \|(Q_0)_{:,k} - (Q_0)_{:,l}\|_2^2.$$

Identifiability. Let us recall that identifiability means that the distribution over all output sequences uniquely determines the model parameters, up to a permutation of its hidden states (see for example Weiss and Nadler [2015]). We consider the following additional assumption.

Assumption A2. $D^2 = \min_{l\neq k} \|(Q_0)_{:,k} - (Q_0)_{:,l}\|_2^2 > 0$.

If Assumptions A1 and A2 hold and if $\alpha_n \log(n) \leq 1/L$, Theorems 2.5, 2.4 and 2.3 in Section 2.2.3 prove that we are able to get consistent estimation of the parameters P, π and Q of our model when

$$\alpha_n = \Omega\left(\frac{\log(n)}{n}\right).$$

Stated otherwise, Assumptions A1, A2 and $\alpha_n \log(n) \leq 1/L$ are sufficient conditions for *learnability* when the average degree of the nodes is of order $\log(n)$ or higher. Learnability is defined as the possi-

bility to estimate consistently the model parameters. Note that learnability implies identifiability. We refer for example to Weiss and Nadler [2015] for further details. Note that the former paper suggests that the condition D^2 may not be necessary for identifiability since in classical Hidden Markov Models, the additional temporal structure allows for identifiability even, say, when some states have exactly the same output distributions.

Error measure. Given two partitions $\hat{G} = (\hat{G}_1, \dots, \hat{G}_K)$ and $G = (G_1, \dots, G_K)$ of $[n]$ into K non-void groups, we define the proportion of non-matching points

$$\mathrm{err}(\hat{G}, G) = \min_{\sigma \in \mathcal{S}_K} \frac{1}{2n} \sum_{k=1}^{K} \left| \hat{G}_k \,\Delta\, G_{\sigma(k)} \right|,$$

where $A \,\Delta\, B = (A \setminus B) \cup (B \setminus A)$ represents the symmetric difference between the two sets A and B, $|A|$ is the cardinality of the set A and $\mathcal{S}_K$ represents the set of permutations on $\{1, \dots, K\}$. When $\hat{G}$ is a partition estimating G, we refer to $\mathrm{err}(\hat{G}, G)$ as the misclassification proportion (or error) of the clustering.

2.2.2 Partial recovery bound for the MSBM

Using the clustering algorithm from Giraud and Verzelen [2019] to infer the hidden communities, we provide a partial recovery bound in the Stochastic Block Model when the communities are assigned through a Markovian dynamic. In the following, $(\hat{C}_i)_{1 \le i \le n}$ and $(\hat{G}_k)_{k \in [K]}$ denote respectively the estimators of $(C_i)_{1 \le i \le n}$ and $(G_k)_{k \in [K]}$ provided by the Algorithm 1 from Giraud and Verzelen [2019] which is described in Section 2.8.
We define the signal-to-noise ratio as

$$S^2 := \frac{n \alpha_n \pi_m D^2}{L},$$

reminding that $\pi_m = \min_{c \in [K]} \pi(c)$, $\|Q_0\|_\infty \le L$ and $D^2 = \min_{l \neq k} \|(Q_0)_{:,k} - (Q_0)_{:,l}\|_2^2$. The SNR should be understood as the ratio between $i)$ the *signal* $\alpha_n^2 n \pi_m D^2$, which is an asymptotic lower bound on the minimal distance between two distinct centers Δ^2 defined as

$$\Delta^2 := \min_{k \neq j} \sum_l |G_l| (Q_{k,l} - Q_{j,l})^2 = \alpha_n^2 \sum_l |G_l| ((Q_0)_{k,l} - (Q_0)_{j,l})^2,$$

and $ii)$ the *noise* $\alpha_n L$. We shed light on the fact that this quantity matches asymptotically the SNR from [Giraud and Verzelen, 2019, Theorem 2] (cf. Section 2.8 for details) when π is the uniform distribution over $[K]$ and when the communities are assigned independently to each node according to the probability distribution π. Moreover, π_m can be related to standard quantities that measure how fast the chain converges to its stationary distribution π. Proposition 2.1 states a direct connection between π_m and the mixing time of the chain.

Proposition 2.1. *[cf. Levin, 2017, Theorem 12.3]*
In the following, we denote $\|\cdot\|_{\mathrm{TV}}$ the total variation norm. Let P be the transition matrix of a reversible, irreducible Markov chain with finite state space E and stationary measure π such that $\pi_m := \min_{x \in E} \pi(x) > 0$. For any $0 < \epsilon < 1$, let

$$t_{\mathrm{mix}}(\epsilon) := \min\{t > 0 \ : \ \sup_x \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \le \epsilon\},$$

be the mixing time of the chain. Then it holds

$$t_{\mathrm{mix}}(\epsilon) \le \log((\epsilon \pi_m)^{-1}) / (1 - \lambda_+),$$

where $1 - \lambda_+$ is the right $\mathcal{L}_2$-spectral gap of the chain from Definition A.12.

The smaller π_m, the slower the convergence of the chain towards π and the smaller the SNR. Similarly to Theorem 2.17, we prove with Theorem 2.2 that the misclassification error decays exponentially fast with respect to the SNR S^2.
In Theorem 2.2, the constants a and b only depend on the parameters π, P and Q_0 while the constant b also depends on the number of communities K. Those constants are made explicit in Lemma 2.12 (cf. Section 2.7).

Theorem 2.2. *Assume that $\alpha_n \log(n) \leq 1/L$. Then there exist three constants $a, b, c > 0$ such that for any n satisfying*

$$n\alpha_n > a,$$

it holds with probability at least $1 - b/n^2$,

$$\mathrm{err}(\hat{G}, G) \leq e^{-cS^2}.$$

In particular, it holds with probability at least $1 - b/n^2$,

$$-\log\Big(\mathrm{err}(\hat{G}, G)\Big) = \Omega(n\alpha_n).$$

Remark: Sparsity and theoretical guarantees. Theorem 2.2 states that in the relatively sparse regime (i.e. when $\alpha_n \sim \log(n)/n$), we achieve a polynomial decay of the misclassification error with order $\pi_m D^2/L$. The greater the quantity $\pi_m D^2/L$ is, the faster the misclassification error decays. In particular, for n large enough it holds with high probability $\mathrm{err}(\hat{G}, G) < 1/(2n)$ which gives $\hat{G} = G$.
The condition on the sparsity parameter α_n indicates that Theorem 2.2 can still be informative in the sparse regime (i.e. when $\alpha_n \sim 1/n$). Typically if $\lim_{n\to\infty} \alpha_n n > A$ for some $A > a$, then Theorem 2.2 ensures that for n large enough it holds with high probability, $\mathrm{err}(\hat{G}, G) \leq e^{-cA\pi_m D^2/L}$.

2.2.3 Consistent estimation of the parameters

Note that Theorem 2.2 is a straightforward consequence of the work of Giraud and Verzelen [2019]. Our methods from Sections 2.3, 2.4 and 2.5 could easily be applied using your favorite clustering algorithm and we have decided to use this recent SDP method for our simulations. In this section, we give estimators $\hat{\pi}$, $\hat{P}$ and $\hat{Q}$ of the parameters of our model, namely π, P and Q. We prove that there are consistent for the infinity norm when the average degree is of order $\log n$ or higher.
In Theorems 2.3, 2.4 and 2.5 the constants a and b' only depend on the parameters π, P and Q_0 while the constant b also depends on the number of communities K. Lemmas 2.14, 2.15 and 2.16 provide respectively a more complete version of Theorems 2.3, 2.4 and 2.5 by giving explicitly these constants.
In Theorems 2.3, 2.4 and 2.5, the condition on the sparsity parameter α_n indicates that we get consistent estimation respectively of the transition matrix, the stationary measure and the connectivity matrix in the relatively sparse regime (i.e. when $\alpha_n \sim \log(n)/n$) for n large enough when $\lim\limits_{n\to\infty} n\alpha_n/\log(n) > a$.

2.2.3.1 The connectivity matrix

In the relatively sparse setting (i.e. when $\alpha_n \sim \log(n)/n$), Theorem 2.2 ensures that for n large enough it holds with high probability $\mathrm{err}(\hat{G}, G) < 1/n$ which implies that the partition of the nodes is correctly recovered. In this case, a natural estimator for $Q_{k,l}$ (for $k, l \in [K]^2$) consists in computing the ratio between the number of edges between nodes with communities k and l and the maximum number of edges between nodes with communities k and l. For any $k, l \in [K]^2$,

$$\hat{Q}_{k,l} := \begin{cases} \dfrac{1}{|\hat{G}_k| \times |\hat{G}_l|} \displaystyle\sum_{i\in\hat{G}_k} \sum_{j\in\hat{G}_l} X_{i,j} & \text{if } k \neq l \\ \dfrac{1}{|\hat{G}_k| \times (|\hat{G}_k| - 1)} \displaystyle\sum_{i,j\in\hat{G}_k} X_{i,j} & \text{if } k = l \end{cases}.$$

One can remark that each entry of our estimator $\hat{Q}$ is a sum of identically distributed and independent Bernoulli random variables (i.e. it is a Binomial random variable). In Section 2.7.2, we prove Theorem 2.3 which ensures the consistency of our estimate of the connectivity matrix.

Theorem 2.3. *Let us consider $\gamma > 0$. Assume that $\alpha_n \log(n) \leq 1/L$. Then there exist three constants $a, b, b' > 0$ such that for any n satisfying*

$$\frac{n\alpha_n}{\log(n)} \geq a \quad \textit{and} \quad n > \left(\frac{\gamma+1}{\pi_m}\right)^2,$$

it holds with probability at least $1 - b(1/n^2 \vee \exp(-b'\gamma^2))$,

$$\|\hat{Q} - Q\|_\infty \leq \frac{\gamma}{\sqrt{n}}.$$

2.2.3.2 The stationary distribution of the Markov chain

Thanks to the ergodic theorem, we know that the average number of visits in each state of the chain converges toward the stationary probability of the chain at this particular state. Stated otherwise, for all community $k \in [K]$, the average number of nodes with community k in the graph converges toward $\pi(k)$ as n tends to $+\infty$. Therefore we propose to estimate the stationary measure of the chain $(C_i)_{i\geq 1}$ with $\hat{\pi}$ defined by

$$\forall k \in [K], \quad \hat{\pi}_k := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{C}_i = k}.$$

Theorem 2.4 ensures the consistency of our estimate $\hat{\pi}$. Its proof can be found in Section 2.7.3.

Theorem 2.4. *Let us consider* $\gamma > 0$.
Assume that $\alpha_n \log(n) \leq 1/L$. *Then there exist three constants* $a, b, b' > 0$ *such that for any* n *satisfying*

$$\frac{n\alpha_n}{\log(n)} \geq a,$$

it holds with probability at least $1 - b(1/n^2 \vee \exp(-b'\gamma^2))$,

$$\|\hat{\pi} - \pi\|_\infty \leq \frac{\gamma}{\sqrt{n}}.$$

2.2.3.3 The transition matrix of the Markov chain

We define $(Y_i)_{i\geq 1}$ a Markov Chain on $[K]^2$ by setting $Y_i = (C_i, C_{i+1})$. We define naturally the sequence $\left(\hat{Y}_i\right)_{i\geq 1}$ by $\hat{Y}_i = (\hat{C}_i, \hat{C}_{i+1})$. The transition kernel of the Markov Chain $(Y_i)_{i\geq 1}$ is $\mathcal{P}_{(k,l),(k',l')} = \mathbb{1}_{l=k'} P_{l,l'}$ and its stationary measure is given by μ such that $\forall k, l$, $\mu(k,l) = \pi(k)P_{k,l}$. We propose to estimate each entry of the transition matrix P of the Markov chain $(C_i)_{i\geq 1}$ with

$$\forall k, l \in [K]^2, \quad \hat{P}_{k,l} := \frac{n}{n-1}\frac{\sum_{i=1}^{n-1} \mathbb{1}_{\hat{Y}_i = (k,l)}}{\sum_{i=1}^{n} \mathbb{1}_{\hat{C}_i = k}}.$$

Theorem 2.5. *Let us consider* $\gamma > \frac{5K}{2\pi_m^2}$. *Assume that* $\alpha_n \log(n) \leq 1/L$. *Then there exist three constants* $a, b, b' > 0$ *such that for any* n *satisfying*

$$\frac{n\alpha_n}{\log(n)} \geq a \quad \textit{and} \quad n\alpha_n \geq \frac{a}{\gamma^2},$$

it holds with probability at least $1 - b\left[1/n^2 \vee \exp\left(-b'(\gamma - \frac{5K}{2\pi_m^2})^2\right)\right]$,

$$\|\hat{P} - P\|_\infty \leq \frac{\gamma}{\sqrt{n}}.$$

Remark. To prove this theorem, we consider the Markov chain $(Y_i)_{i\geq 1}$ built considering two consecutive states of the Markov chain $(C_i)_{i\geq 1}$. Stated otherwise, the state number i of the Markov chain used is formed by the couple of the communities of the nodes number i and number $i+1$.

Notation. In the following sections, $\hat{\mathbb{P}}$ will denote the probability measure under which the Markov chain $(C_i)_{i\geq 1}$ has transition matrix $\hat{P}$ and $X_{i,j} \sim \mathrm{Ber}(\hat{Q}_{C_i,C_j})$ for all $i, j \in [n],\ i \neq j$.

2.3 Markovian dynamic testing

We illustrate our model on a toy example with $K = 4$ communities, with the transition matrix P and the connectivity matrix Q defined by

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.5 & 0.1 \\ 0.45 & 0.15 & 0.2 & 0.2 \\ 0.15 & 0.3 & 0.1 & 0.45 \\ 0.25 & 0.3 & 0.1 & 0.35 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} 0.22 & 0.48 & 0.29 & 0.44 \\ 0.48 & 0.61 & 0.18 & 0.15 \\ 0.29 & 0.18 & 0.08 & 0.87 \\ 0.44 & 0.15 & 0.87 & 0.27 \end{bmatrix}. \tag{2.1}$$

All the experiments presented in our paper can be reproduced using the Python notebooks provided on this repository[1].

As a first application of our model, we propose a hypothesis test to statistically distinguish between an independent assignment of the communities with the distribution π and a Markovian assignment with a non-trivial dependence structure. More precisely, we consider the null $\mathbb{H}_0$: *communities are independently assigned with distribution* π where π denotes the stationary distribution of the transition matrix P from (2.1). Our test is based on estimate $\hat{P}$ of the transition matrix. The null can be rephrased as $\mathbb{H}_0 : \ P = P^0$ where $P^0 := \begin{bmatrix} \pi \\ \vdots \\ \pi \end{bmatrix}$. One can use any *black-box goodness-of-fit test* comparing $\hat{P}$ to P^0.

Figure 2.2 shows the power of this hypothesis test with level 5% (Type I error) and using the χ^2-test described by [Bickenbach et al., 2001, Section 2.4]. Rejection region is calibrated (i.e., threshold of the χ^2-test) by *Monte Carlo simulations under the null*. It allows us to control Type I error as depicted by dotted blue line. We run our algorithm to estimate the transition matrix from which we compute the χ^2-test statistic namely

$$S_n := \sum_{1 \le k,l \le K} |\hat{G}_k| \frac{\left(\hat{P}_{k,l} - \pi_l\right)^2}{\pi_l} \text{ with } |\hat{G}_k| = \sum_{i=1}^{n} \mathbb{1}_{\hat{C}_i = k}.$$

S_n is known to be asymptotically distributed as a χ^2 random variable with $K(K-1)$ degrees of freedom.

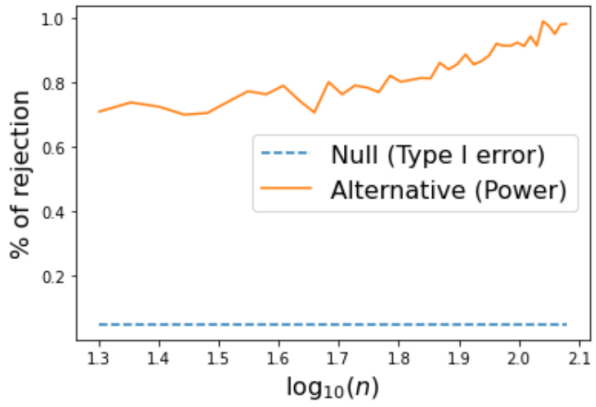


Figure 2.2: Power of our hypothesis test with level 5%. We choose alternative given by the matrices defined in (2.1). We see that for graphs of size larger than 100, the rejection rate is almost 1 under the alternative (Type II error is almost zero), the test is very powerful.

[1] https://github.com/quentin-duchemin/inference-markovian-SBM

2.4 Link prediction

2.4.1 The plug-in approach

We propose to show the usefulness of our model solving a link prediction problem taking into account the underlying time dynamic, which cannot be done with classical SBMs (iid framework). Link prediction consists in forecasting how future nodes will connect to the rest of the graph. Considering a graph of size n sampled from MSBM, link prediction can be achieved using a *forward step* on our Markovian dynamic, giving the posterior probability of Definition 2.6.

Definition 2.6. (Posterior probability function)
Let us consider a graph X of $n+1$ nodes generated from the model described in Section 2.2.1. We consider $\mathbf{c}_{1:n}$ a sequence of communities for the n first nodes. Then the posterior probability function η is defined by

$$\forall i \in [n], \quad \eta_i(\mathbf{c}_{1:n}) = \mathbb{P}\left(X_{i,n+1} = 1 \mid \mathbf{C}_{1:n} = \mathbf{c}_{1:n}\right).$$

We consider a classifier g (see Definition 2.7) and an algorithm that, given some communities $\mathbf{c}_{1:n}$, estimates $X_{i,n+1}$ by putting an edge between nodes i and $n+1$ if $g_i(\mathbf{c}_{1:n})$ is 1.

Definition 2.7. A *classifier* is a function which associates to any sequence of communities $\mathbf{c}_{1:n}$ binary variables $(g_i(\mathbf{c}_{1:n}))_{i\in[n]} \in \{0,1\}^n$.

The risk of this algorithm is as in *binary classification*,

$$\begin{aligned}\mathcal{R}(g, \mathbf{c}_{1:n}) :=& \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(g_i(\mathbf{C}_{1:n}) \neq X_{i,n+1} \mid \mathbf{C}_{1:n} = \mathbf{c}_{1:n}\right) \\ =& \frac{1}{n}\sum_{i=1}^{n} (1-\eta_i(\mathbf{c}_{1:n}))\mathbb{1}_{g_i(\mathbf{c}_{1:n})=1} + \eta_i(\mathbf{c}_{1:n})\mathbb{1}_{g_i(\mathbf{c}_{1:n})=0},\end{aligned} \tag{2.2}$$

Pushing further this analogy, we can define the classification error of some classifier g by $L(g) = \mathbb{E}\left[\mathcal{R}(g, \mathbf{C}_{1:n})\right]$. Proposition 2.9 shows that the Bayes estimator - introduced in Definition 2.8 - is optimal for the risk defined in (2.2).

Definition 2.8. (Bayes estimator)
We keep the notations of Definition 2.6. The Bayes estimator g^* of $(X_{i,n+1})_{1\leq i\leq n}$ is defined by

$$\forall i \in [n], \quad g_i^*(\mathbf{c}_{1:n}) = \begin{cases} 1 & \text{if } \eta_i(\mathbf{c}_{1:n}) \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 2.9. *(Optimality of the Bayes classifier for the risk $\mathcal{R}$)*
We keep the notations of Definitions 2.6 and 2.8. For any classifier g, it holds for any $i \in [n]$,

$$\begin{aligned}&\mathbb{P}\left(g_i(\mathbf{C}_{1:n}) \neq X_{i,n+1} \mid \mathbf{C}_{1:n} = \mathbf{c}_{1:n}\right) - \mathbb{P}\left(g_i^*(\mathbf{C}_{1:n}) \neq X_{i,n+1} \mid \mathbf{C}_{1:n} = \mathbf{c}_{1:n}\right) \\ &= 2\left|\eta_i(\mathbf{c}_{1:n}) - \frac{1}{2}\right| \times \mathbb{E}\left\{\mathbb{1}_{g_i(\mathbf{C}_{1:n})\neq g_i^*(\mathbf{C}_{1:n})} \mid \mathbf{C}_{1:n} = \mathbf{c}_{1:n}\right\},\end{aligned}$$

which immediately implies that

$$\mathcal{R}(g, \mathbf{c}_{1:n}) \geq \mathcal{R}(g^*, \mathbf{c}_{1:n}) \text{ and therefore } L(g) \geq L(g^*).$$

A natural question arises: Can we approximate the Bayes classifier given the observation of the adjacency matrix $X \in \mathbb{R}^{n\times n}$? A reasonable candidate is the MSBM classifier introduced in Definition 2.10.

Definition 2.10. (The MSBM classifier) For any n and any $i \in [n]$, we define

$$\hat{\eta}_i(\hat{\mathbf{C}}_{1:n}) = \sum_{k\in[K]} \hat{Q}_{\hat{C}_i,k}\hat{P}_{\hat{C}_n,k}, \tag{2.3}$$

where $\hat{Q}$ and $\hat{P}$ denote respectively the estimate of the connections matrix and the transition matrix

with our method (see Section 2.2.3). The MSBM classifier is defined by

$$\forall i \in [n], \quad g_i^{MSBM}(\hat{\mathbf{C}}_{1:n}) = \begin{cases} 1 & \text{if } \hat{\eta}_i(\hat{\mathbf{C}}_{1:n}) \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 2.11 shows that the MSBM classifier is consistent, meaning that given a training set, the probability of correct classification approaches - as the size of the training set increases - the best probability theoretically possible if the population distributions were fully known.

Proposition 2.11. *(Consistency of the MSBM classifier) Let us consider* $\gamma > \frac{5K}{2\pi_m^2}$. *Assume that* $\alpha_n \log(n) \leq 1/L$. *Then there exist three constants* $a, b, b' > 0$ *such that for any* n *satisfying*

$$\frac{n\alpha_n}{\log(n)} \geq a, \; n\alpha_n \geq \frac{a}{\gamma^2} \;\; \textit{and} \;\; n > \left(\frac{\gamma+1}{\pi_m}\right)^2, \tag{2.4}$$

it holds with probability at least $1 - b\left[1/n \vee n\exp\left(-b'(\gamma - \frac{5K}{2\pi_m^2})^2\right)\right]$,

$$\forall i \in [n], \quad |\eta_i(\mathbf{C}_{1:n}) - \hat{\eta}_i(\mathbf{C}_{1:n})| \leq \frac{\gamma}{\sqrt{n}}\left(\alpha_n KL + 1\right). \tag{2.5}$$

Using Theorem 2.2, we deduce that for n *large enough, (2.5) holds replacing* $\hat{\eta}_i(\mathbf{C}_{1:n})$ *by* $\hat{\eta}_i(\hat{\mathbf{C}}_{1:n})$.

Remark. Let us point out that obtaining a non-trivial result from Proposition 2.11 may require to choose γ as function of n. Typically, choosing $\gamma = n^{1/4}$, we obtain that for n large enough, it holds with probability at least $1 - b/n$,

$$\forall i \in [n], \quad |\eta_i(\mathbf{C}_{1:n}) - \hat{\eta}_i(\mathbf{C}_{1:n})| \leq n^{-1/4}\left(\alpha_n KL + 1\right).$$

Figure 2.3 illustrates Proposition 2.11 on graphs with 180 nodes sampled from the MSBM.

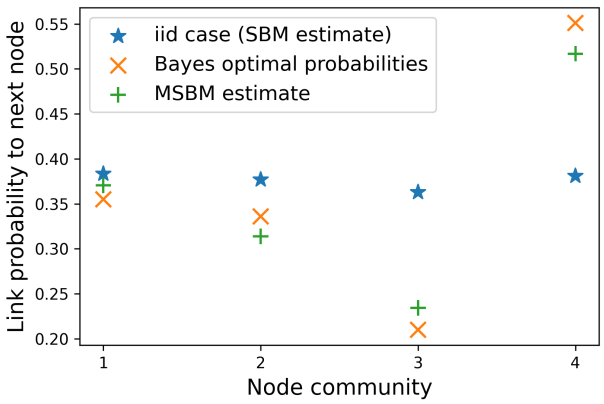


Figure 2.3: We consider $K = 4$ communities and we sample a random graph of size $n = 180$ from the MSBM using matrices defined in (2.1). We plot the averaged link probabilities $|G_c|^{-1}\sum_{i\in G_c}\sum_{k\in[K]}\hat{Q}_{c,k}\hat{P}_{C_{n-1},k}$ for all $c \in [K]$ with $G_c = \{i \in [n-1] \mid C_i = c\}$.

2.4.2 Reliable link prediction

One can notice that despite the nice theoretical property of the MSBM classifier from Definition 2.10, the practical results of such approach can be discussed. Indeed, we only use the two estimates $\hat{C}_i$ and $\hat{C}_n$ to build $\hat{\eta}_i(\hat{C}_{1:n})$ without taking advantage of the complete sequence of recovered communities. To cope with this issue, we propose to learn the probability that a node i belonging to community $k \in [K]$ is assigned to a cluster $l \in [K]$ by the algorithm. These quantities are the emission probabilities of a Hidden Markov Model (HMM) with hidden states the true assignment of the nodes while the

observations are the communities estimated by the algorithm, namely

$$\forall k, l \in [K], \quad O_{k,l} := \mathbb{P}(\hat{C}_1 = l | C_1 = k).$$

The Baum-Welch algorithm allows us to estimate the emission probabilities by performing (i) a forward procedure that learns the probability of seeing the observations $\hat{\mathbf{c}}_{1:i} \in [K]^i$ and being in state k at time i, namely

$$\alpha_k(i) = \mathbb{P}(\hat{\mathbf{C}}_{1:i} = \hat{\mathbf{c}}_{1:i}, C_i = k),$$

and (ii) a backward procedure that learns the probability of the ending partial sequence $\hat{\mathbf{c}}_{j+1:n}$ given being at state l at time j, namely

$$\beta_l(j) = \mathbb{P}(\hat{\mathbf{C}}_{j+1:n} = \hat{\mathbf{c}}_{j+1:n} | C_j = l).$$

Using this approach and recalling the notation $\hat{\mathbb{P}}$ introduced at the end of Section 2.2.3, we can compute

$$\hat{\zeta}_{k,l}^{(i,j)}(\hat{\mathbf{c}}_{1:n}) = \hat{\mathbb{P}}(C_i = k, C_j = l | \hat{\mathbf{C}}_{1:n} = \hat{\mathbf{c}}_{1:n})$$

which estimates

$$\zeta_{k,l}^{(i,j)}(\hat{\mathbf{c}}_{1:n}) := \mathbb{P}(C_i = k, C_j = l | \hat{\mathbf{C}}_{1:n} = \hat{\mathbf{c}}_{1:n}) \propto \alpha_k(i) \chi_{k,l}^{(i,j)} \beta_l(j),$$

where

$$\begin{aligned}\chi_{k,l}^{(i,j)} &:= \mathbb{P}(C_j = l, \hat{\mathbf{C}}_{i+1:j} = \hat{\mathbf{c}}_{i+1:j} | C_i = k) \\ &= \sum_{c_{i+1}, \dots, c_{j-1}} P_{k,c_{i+1}} \left(O_{c_{i+1}, \hat{c}_{i+1}} + P_{c_{i+1}, c_{i+2}} (O_{c_{i+2}, \hat{c}_{i+2}} + \cdots + P_{c_{j-1}, l} O_{l, \hat{c}_j}) \right).\end{aligned}$$

We can then build the Reliable MSBM (RMSBM) classifier by considering

$$\hat{\eta}_i^R(\hat{\mathbf{c}}_{1:n}) = \sum_{k, c_i, c_n \in [K]} \hat{\zeta}_{c_i, c_n}^{(i,n)}(\hat{\mathbf{c}}_{1:n}) \hat{Q}_{c_i, k} \hat{P}_{c_n, k},$$

and then replacing $\hat{\eta}_i$ by $\hat{\eta}_i^R$ in the definition of the MSBM classifier. Note that this approach is a heuristic because of the local optimum reached by the EM algorithm and also because of the dependence of the emission probabilities in our model. Figure 2.4 shows that this method leads to a reliable estimation of the posterior probabilities. The RMSBM classifier gives smaller L^1 errors on the posterior probabilities. The difference is significant when the clustering algorithm fails to recover the complete partition of the nodes which leads to bad estimates for the plug-in approach.

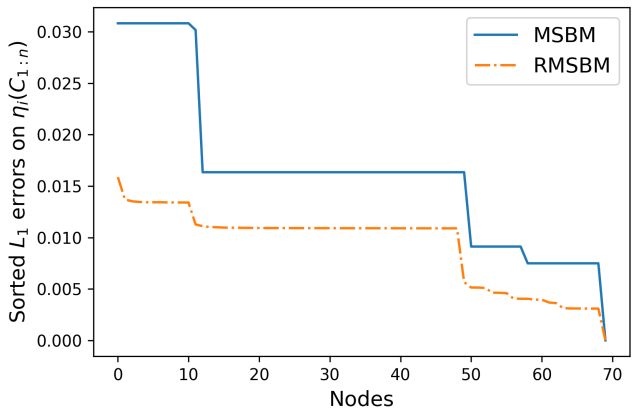


Figure 2.4: We plot the sorted L^1 errors between $\hat{\eta}_i(\hat{\mathbf{C}}_{1:n})$ (resp. $\hat{\eta}_i^R(\hat{\mathbf{C}}_{1:n})$) and $\eta_i(\hat{\mathbf{C}}_{1:n})$. Our reliable estimation of the posterior probabilities allows to get a significantly smaller variance compared to the plug-in approach.

Note that using the Baum-Welch algorithm, we recover the emission probabilities $O_{k,l}$ for all $k, l \in [K]$ and Figure 2.5 shows that they can be used to extract relevant information on the clustering algorithm.

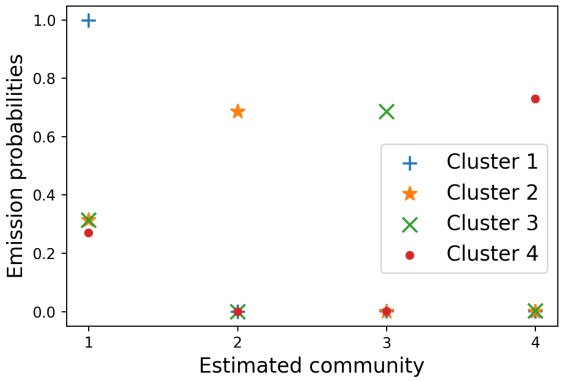


Figure 2.5: We work with a graph of size 120 and with matrices defined by (2.1). We plot the learned emission probabilities $O_{k,l}$, $k, l \in [K]$. The ergodic theorem ensures that the first cluster is (asymptotically) the smaller. Indeed, the stationary measure of P from (2.1) is approximately $[0.14\ ,\ 0.22\ ,\ 0.38\ ,\ 0.26]$. We observe that the errors made by the algorithm consist in assigning nodes from community $2, 3$ or 4 to cluster 1. This means that the clustering algorithm from Giraud and Verzelen [2019] tends to overestimate the size of small clusters.

2.4.3 The Baum-Welch algorithm

In the previous section, we have presented a reliable approach to solve link prediction or a collaborative filtering problem when we fully observe the graph at time n and when we want to perform some temporal prediction involving future nodes. We propose to consider a more general framework considering that we fully observe the graph at time $n + \delta$ ($\delta \in \mathbb{N}^*$) but we consider that edges involving nodes between time T (with $T < n$) and time n are not reliable. Note that the simpler framework addressed in the paper is simply recovered by taking $n = T + 1$. Hence, we want only to take into account the edges involving pairs of nodes in $\{1, \dots, T, n, \dots n + \delta\}$. We denote $E_{T,n,\delta}$ this set of edges. We describe the Baum-Welch algorithm in this framework. Running the clustering algorithm on the graph $G = (\{1, \dots, T, n, \dots n + \delta\}, E_{T,n,\delta})$, we find a sequence of estimates for the communities $\hat{\mathbf{C}}_{1:T}, \hat{\mathbf{C}}_{n:n+\delta}$. In the following, we will consider by abuse of notations that for any $j \geq T + 1$, the sequence $\hat{\mathbf{C}}_{1:j}$ represents the sequence $\left(\hat{\mathbf{C}}_i, i \in [j] \backslash \{T + 1, \dots, n - 1\}\right)$.
The Baum-Welch algorithm consists in a forward and a backward procedure followed by an update step that we describe below. In the following, $\theta = (\tilde{P}, O, \mu)$ will denote the HMM with transition kernel $\tilde{P}$ for the Markov chain $(C_i)_{i \geq 1}$ with initial distribution μ and with matrix of emission probabilities O. Denoting $\mathbf{1}_K = (1, 1, \dots, 1)^\top \in \mathbb{R}^K$, $\theta = (\tilde{P}, O, \mu)$ is initialized as follows

$$\begin{aligned}
\tilde{P} &= \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^\top, \\
\mu &= \mathbf{1}_K^\top, \\
O &= (1 - \epsilon) \mathrm{Id}_K + \frac{\epsilon}{K - 1} \left(\mathbf{1}_K \mathbf{1}_K^\top - \mathrm{Id}_K \right),
\end{aligned}$$

where $\epsilon \in (0, 1)$ (typically $\epsilon = 10^{-2}$).

- Forward procedure

  Let us recall that we have denoted $\alpha_k(i) = \mathbb{P}\left(\hat{\mathbf{C}}_{1:i} = \hat{\mathbf{c}}_{1:i}, C_i = k \mid \theta\right)$ the probability of seeing the

observations $\hat{c}_1, \dots \hat{c}_i$ and being in state k at time i. This is found recursively with

$$\begin{aligned} \forall k \in [K], \quad \alpha_k(1) =& \mu_k O_{k,\hat{c}_1} \\ \forall k \in [K],\ \forall i \in [n], \quad \alpha_k(i) =& \begin{cases} \sum_{l\in[K]} \alpha_l(T) \left(\tilde{P}^{i-T}\right)_{l,k} & \text{if } T < i \leq n \\ O_{k,\hat{c}_i} \sum_{l\in[K]} \alpha_l(i-1)\tilde{P}_{l,k} & \text{otherwise.} \end{cases} \end{aligned}$$

- Backward procedure

  Let us recall that we have denoted $\beta_k(i) = \mathbb{P}\left(\hat{\mathbf{C}}_{i+1:n+\delta} = \hat{\mathbf{c}}_{i+1:n+\delta} \mid C_i = k, \theta\right)$ the probability of the ending partial sequence $\hat{\mathbf{c}}_{i+1:n+\delta}$ given starting in state k at time i. This is found recursively with

$$\begin{aligned} \forall k \in [K], \quad \beta_k(n) =& 1 \\ \forall k \in [K],\ \forall i \in [n], \quad \beta_k(i) =& \begin{cases} \sum_{l\in[K]} \beta_l(n-1) \left(\tilde{P}^{n-1-i}\right)_{k,l} & \text{if } T \leq i \leq n-2 \\ \sum_{l\in[K]} \beta_l(i+1)\tilde{P}_{k,l} O_{l,\hat{c}_{i+1}} & \text{otherwise.} \end{cases} \end{aligned}$$

- Update step

  We can first update the temporary variables γ and ξ defined below. The probability of being in state k at time i given the observed sequence $\hat{\mathbf{C}}_{1:n+\delta} = \hat{\mathbf{c}}_{1:n+\delta}$ and the parameters θ is denoted $\gamma_k(i)$ with

$$\forall k \in [K],\ \forall i \in [n], \quad \gamma_k(i) = \mathbb{P}(C_i = k | \hat{\mathbf{C}}_{1:n+\delta} = \hat{\mathbf{c}}_{1:n+\delta}, \theta) = \frac{\alpha_k(i)\beta_k(i)}{\sum_{l\in[K]} \alpha_l(i)\beta_l(i)}.$$

  The probability of being in state k and l at times i and $i+1$ respectively given the observed sequence $\hat{\mathbf{C}}_{1:n+\delta}$ and parameters θ is denote $\xi_{k,l}(i)$ with for all $k, l \in [K]$ and for all $i \in [n]$,

$$\xi_{k,l}(i) = \mathbb{P}(C_i = k, C_{i+1} = l \mid \hat{\mathbf{C}}_{1:n+\delta} = \hat{\mathbf{c}}_{1:n+\delta}, \theta) = \frac{\mathbb{P}(C_i = k, C_{i+1} = l, \hat{\mathbf{C}}_{1:n+\delta} = \hat{\mathbf{c}}_{1:n+\delta} \mid \theta)}{\mathbb{P}(\hat{\mathbf{C}}_{1:n+\delta} = \hat{\mathbf{c}}_{1:n+\delta} \mid \theta)}.$$

  Hence,

$$\begin{aligned} \xi_{k,l}(i) \quad &= \frac{\alpha_k(i)\tilde{P}_{k,l}\beta_l(i+1)}{\sum_{c,b\in[K]} \alpha_c(i)\tilde{P}_{c,b}\beta_b(i+1)} && \text{if } T \leq i \leq n-2 \\ \xi_{k,l}(i) \quad &= \frac{\alpha_k(i)\tilde{P}_{k,l}\beta_l(i+1)O_{l,\hat{c}_{i+1}}}{\sum_{c,b\in[K]} \alpha_c(i)\tilde{P}_{c,b}\beta_b(i+1)O_{b,\hat{c}_{i+1}}} && \text{otherwise.} \end{aligned}$$

  The parameters of the hidden Markov model θ can now be updated.

$$\begin{aligned} \forall k \in [K], \quad \mu_k &= \gamma_k(1) \\ \forall k, l \in [K], \quad \tilde{P}_{k,l} &= \frac{\sum_{i=1}^{n-1} \xi_{k,l}(i)}{\sum_{i=1}^{n-1} \gamma_k(i)} \\ \forall k, l \in [K], \quad O_{k,l} &= \frac{\sum_{i=1}^{n} \mathbb{1}_{\hat{c}_i = l}\gamma_k(i)}{\sum_{i=1}^{n} \gamma_k(i)}. \end{aligned}$$

2.5 Collaborative filtering

2.5.1 Reliable collaborative filtering

Let us now dig into another prediction question, namely collaborative filtering. Solving a collaborative filtering task consists in inferring the community of one node of the graph if we have only partial information on how this node connects to the rest of the graph. More precisely, we observe fully the graph at time m and for some $n > m$, we observe how the node n is connected (or not) to a subset of nodes $\mathcal{E} \subset [m]$, i.e. we have access to $(X_{i,n})_{i\in\mathcal{E}}$. Our goal is then to predict the community of node n:

C_n. We propose to use the maximum a posteriori (MAP) estimator to tackle this problem. The optimal MAP selects

$$\hat{C}_n \in \underset{k\in[K]}{\arg\max}\, \mathbb{P}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \mathbf{C}_{1:m}\right),$$

while given a sequence of estimated communities $\hat{\mathbf{c}}_{1:m}$, the plug-in MAP selects

$$\hat{C}_n^{PI} \in \underset{k\in[K]}{\arg\max}\, \hat{\mathbb{P}}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \mathbf{C}_{1:m} = \hat{\mathbf{c}}_{1:m}\right),$$

and the Reliable MAP chooses

$$\hat{C}_n^{R} \in \underset{k\in[K]}{\arg\max}\, \hat{\mathbb{P}}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m}\right).$$

Denoting $\mathcal{E} = \{i_1, \dots, i_S\}$ with $1 \le i_1 < \cdots < i_S \le m$, $\hat{C}_n^R$ can be computed noticing that

$$\begin{aligned}
&\underset{k\in[K]}{\arg\max}\, \hat{\mathbb{P}}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m}\right) \\
&= \underset{k\in[K]}{\arg\max}\, \hat{\mathbb{P}}\left((X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m} \mid C_n = k\right) \times \hat{\mathbb{P}}\left(C_n = k\right),
\end{aligned}$$

with

$$\begin{aligned}
&\hat{\mathbb{P}}\left((X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m} \mid C_n = k\right) \\
&= \sum_{c_{i_1},\dots,c_{i_S}\in[K]} \alpha_{c_{i_1}}(i_1) \prod_{j=1}^{S-1} \left(\hat{l}_{i_j,n}(c_{i_j},k)\hat{\chi}^{(i_j,i_{j+1})}_{c_{i_j},c_{i_{j+1}}}\right) \beta_{c_{i_S}}(i_S)\hat{l}_{i_S,n}(c_{i_S},k),
\end{aligned}$$

where $\hat{l}_{i,n}(c,k) = \hat{Q}_{c,k}^{X_{i,n}}(1-\hat{Q}_{c,k})^{1-X_{i,n}}$ and $\hat{\mathbb{P}}\left(C_n = k\right) = \left(\mu\hat{P}^n\right)_k$. In the last equality, μ is the initial distribution of the Markov chain $(X_i)_{i\ge1}$ which is learned by the Baum-Welch algorithm (see Section 2.4.3). Figure 2.6 compares on a numerical experiment the plug-in MAP and the Reliable MAP. The plug-in MAP gives reasonable results but its misclassification rate is always lower bounded by the one of the Reliable MAP.

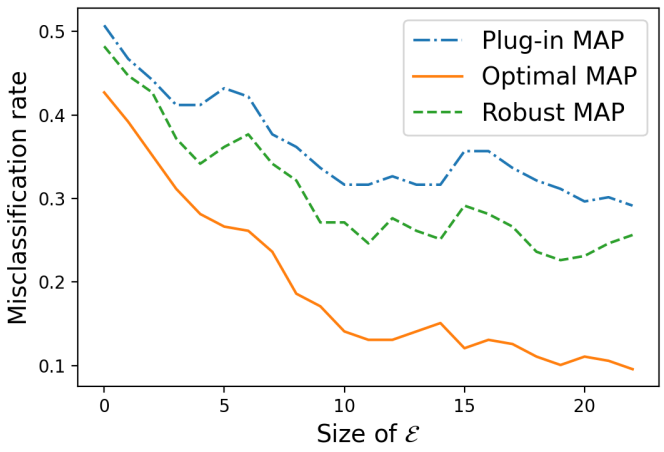


Figure 2.6: We consider a random graph drawn from the MSBM using the matrices given in (2.1). We fully observe the graph until time $m = 100$ and we observe how the node $n = 120$ is connected to the nodes in $\mathcal{E}$ where $\mathcal{E}$ is equal to $\{m\}$, $\{m-1, m\}$, ... or $\{m-25, \dots, m\}$. For those different choices of $\mathcal{E}$, we plot the average error on the clustering of the node n using the optimal MAP, the plug-in MAP or the Reliable MAP.

2.5.2 Theoretical justifications

In this section, we simply derive the formula to compute the different estimates of the community of node n in the collaborative filtering problem tackled in the previous section. In the following, the symbol $\propto$ will be used in the sense that the considered quantities are equal up to a constant that does not depend on the community $C_n = k$ of the node n.

- The optimal MAP selects $\hat{C}_n \in \underset{k\in[K]}{\arg\max}\, \mathbb{P}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \mathbf{C}_{1:m}\right)$ with

$$\begin{aligned}\mathbb{P}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \mathbf{C}_{1:m}\right) &\propto \mathbb{P}\left(C_n = k, (X_{i,n})_{i\in\mathcal{E}} \mid \mathbf{C}_{1:m}\right)\\ &= \mathbb{P}\left((X_{i,n})_{i\in\mathcal{E}} \mid C_n = k, \mathbf{C}_{1:m}\right) \times \mathbb{P}\left(C_n = k \mid \mathbf{C}_{1:m}\right)\\ &= \prod_{i\in\mathcal{E}} Q_{C_i,k}^{X_{i,n}} (1 - Q_{C_i,k})^{X_{i,n}} \times \left(P^{n-m}\right)_{C_m,k}.\end{aligned}$$

- The plug-in MAP selects $\hat{C}_n^{PI} \in \underset{k\in[K]}{\arg\max}\, \hat{\mathbb{P}}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \mathbf{C}_{1:m} = \hat{\mathbf{c}}_{1:m}\right)$ with

$$\begin{aligned}&\hat{\mathbb{P}}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \mathbf{C}_{1:m} = \hat{\mathbf{c}}_{1:m}\right)\\ &\propto \hat{\mathbb{P}}\left(C_n = k, (X_{i,n})_{i\in\mathcal{E}} \mid \mathbf{C}_{1:m} = \hat{\mathbf{c}}_{1:m}\right)\\ &= \hat{\mathbb{P}}\left((X_{i,n})_{i\in\mathcal{E}} \mid C_n = k, \mathbf{C}_{1:m} = \hat{\mathbf{c}}_{1:m}\right) \times \hat{\mathbb{P}}\left(C_n = k \mid \mathbf{C}_{1:m} = \hat{\mathbf{c}}_{1:m}\right)\\ &= \prod_{i\in\mathcal{E}} \hat{Q}_{\hat{c}_i,k}^{X_{i,n}} (1 - \hat{Q}_{\hat{c}_i,k})^{X_{i,n}} \times \left(\hat{P}^{n-m}\right)_{\hat{c}_m,k}.\end{aligned}$$

- The Reliable MAP selects $\hat{C}_n^{R} \in \underset{k\in[K]}{\arg\max}\, \hat{\mathbb{P}}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m}\right)$ with

$$\begin{aligned}\hat{\mathbb{P}}\left(C_n = k \mid (X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m}\right) &\propto \hat{\mathbb{P}}\left(C_n = k, (X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m}\right)\\ &= \hat{\mathbb{P}}\left((X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m} \mid C_n = k\right) \times \hat{\mathbb{P}}\left(C_n = k\right).\end{aligned}$$

We have easily $\hat{\mathbb{P}}\left(C_n = k\right) = \sum_{l\in[K]} \mu_l \left(\hat{P}^n\right)_{l,k}$. Moreover,

$$\begin{aligned}&\hat{\mathbb{P}}\left((X_{i,n})_{i\in\mathcal{E}}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m} \mid C_n = k\right)\\ &= \hat{\mathbb{P}}\left((X_{i_j,n})_{j\in[S]}, \hat{\mathbf{C}}_{1:m} = \hat{\mathbf{c}}_{1:m} \mid C_n = k\right)\\ &= \sum_{c_{i_1}\in[K]} \alpha_{c_{i_1}}(i_1) \hat{Q}_{c_{i_1},k}^{X_{i_1,n}} (1 - \hat{Q}_{c_{i_1},k})^{X_{i_1,n}}\\ &\qquad\qquad \times \hat{\mathbb{P}}\left((X_{i_j,n})_{j\in\{2,\dots,S\}}, \hat{\mathbf{C}}_{i_1+1:m} = \hat{\mathbf{c}}_{i_1+1:m} \mid C_{i_1} = c_{i_1}, C_n = k\right)\\ &= \sum_{c_{i_1}, c_{i_2}\in[K]} \alpha_{c_{i_1}}(i_1) \hat{l}_{i_1,n}(c_{i_1}, k) \underbrace{\hat{\mathbb{P}}\left(\hat{\mathbf{C}}_{i_1+1:i_2} = \hat{\mathbf{c}}_{i_1+1:i_2}, C_{i_2} = c_{i_2} \mid C_{i_1} = c_{i_1}\right)}_{=\hat{\chi}_{c_{i_1},c_{i_2}}^{(i_1,i_2)}} \hat{l}_{i_2,n}(c_{i_2}, k)\\ &\qquad \times \hat{\mathbb{P}}\left((X_{i_j,n})_{j\in\{3,\dots,S\}}, \hat{\mathbf{C}}_{i_2+1:m} \mid C_{i_2} = c_{i_2}, C_n = k\right)\\ &= \dots\\ &= \sum_{c_{i_1},\dots,c_{i_S}\in[K]} \alpha_{c_{i_1}}(i_1) \prod_{j=1}^{S-1} \left(\hat{l}_{i_j,n}(c_{i_j}, k) \hat{\chi}_{c_{i_j},c_{i_{j+1}}}^{(i_j,i_{j+1})}\right) \beta_{c_{i_S}}(i_S) \hat{l}_{i_S,n}(c_{i_S}, k),\end{aligned}$$

where we have denoted $\forall i,j,\ \forall k,l \in [K]$,

$$\hat{\chi}_{k,l}^{(i,j)} = \hat{\mathbb{P}}\left(C_j = l, \hat{\mathbf{C}}_{i+1:j} = \hat{\mathbf{c}}_{i+1:j} |\ C_i = k\right)$$
$$= \sum_{c_{i+1},\dots,c_{j-1}} \hat{P}_{k,c_{i+1}} \left(O_{c_{i+1},\hat{c}_{i+1}} + \hat{P}_{c_{i+1},c_{i+2}} (O_{c_{i+2},\hat{c}_{i+2}} + \cdots + \hat{P}_{c_{j-1},l} O_{l,\hat{c}_j}) \right),$$

and $\forall i,\ \forall c,k \in [K], \quad \hat{l}_{i,n}(c,k) = \hat{Q}_{c,k}^{X_{i,n}} (1-\hat{Q}_{c,k})^{1-X_{i,n}}$.

2.6 Implementation and Experiments

2.6.1 Performance and implementation

The clustering algorithm used is a SDP method and, as a consequence, its time complexity scales with n^3, while the complexity of the Baum-Welch algorithm is of order $K^2 n$. From here, computing $\hat{\eta}_i(\hat{\mathbf{c}}_{1:n})$ for all $i \in [n]$ requires $K^3 n^2$ operations using dynamic programming (see Sec.3.b of the notebook *experiments.ipynb*). Regarding the collaborative filtering task, using again dynamic programming the Reliable MAP estimator from Section 2.5.1 has a time complexity of order $K^4 n^2$ (see method *collaborative_filtering_robustMAP* in the file *markovianSBM/BaumWelch.py*).

2.6.2 Inferring the number of communities

In this section, we propose a heuristic based on the learned emission probabilities from the Baum-Welch algorithm to estimate the number of communities K of our model. The proposed approach consists in running the Baum-Welch algorithm for a finite list of possible number of clusters $\{K_{min},\dots,K_{max}\} = \mathcal{K} \subset \mathbb{N}^*$. For each $K \in \mathcal{K}$, we denote $O^{(K)}$ the matrix of emission probabilities learned by the algorithm when we consider that the number of communities is K. For any $K \in \mathcal{K}$, we define

$$M^{(K)} := \max_{k,l \in [K], k \neq l} \left\{ O_{l,k}^{(K)} + O_{k,l}^{(K)} \right\}.$$

For any $K \in \mathcal{K}$ and any $k,l \in [K],\ k \neq l$, $O_{l,k}^{(K)} + O_{k,l}^{(K)}$ represents the probability that the clustering algorithm predicts community k or l if the true cluster is the other one.

- When K is less than or equal to the true number of clusters, $M^{(K)}$ stays small as soon as the graph is large enough and as the clustering algorithm used is efficient.
- When K becomes greater then the true number of clusters, $M^{(K)}$ is larger compared to the previous case because at least one true cluster will be arbitrarily split in two different groups by the clustering algorithm.

Based on this remark, we propose to estimate the number of communities by choosing the value $K \in \mathcal{K}$ leading to the larger positive jump of the function $K \mapsto M^{(K)}$ namely

$$\hat{K} \in \underset{K \in \{K_{min},\dots,K_{max}-1\}}{\arg\max} \left\{ M^{(K+1)} - M^{(K)} \right\}.$$

First we test our method with a graph of size $n = 110$ and with $K = 4$ communities using the transition kernel P and the connectivity matrix Q defined by (2.1). Figure 2.7 shows that our approach allows to estimate the correct number of communities $K = 4$.
We also test our procedure on a real network corresponding to American football games between Division IA colleges during regular season Fall 2000. Two teams are connected if they played against each other. The nodes have values that indicate which conferences the corresponding team belongs to. We worked with 6 different conferences [2]. Figure 2.8 shows that our procedure infers the correct number of communities.

[2] namely Atlantic Coast, Big East, Big Ten, Big Twelve, Conference USA, Mid-American.

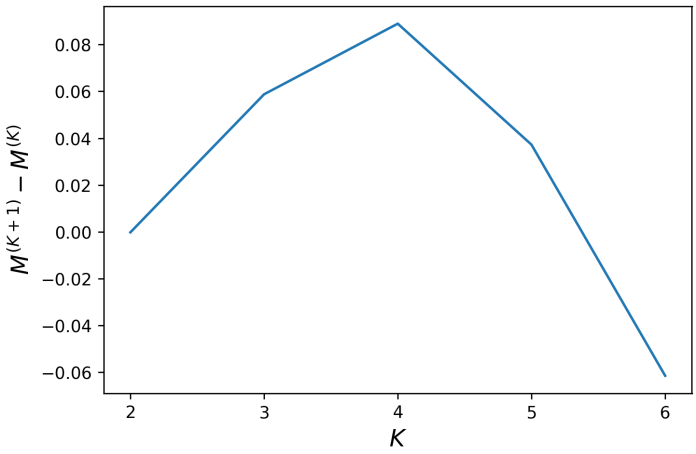
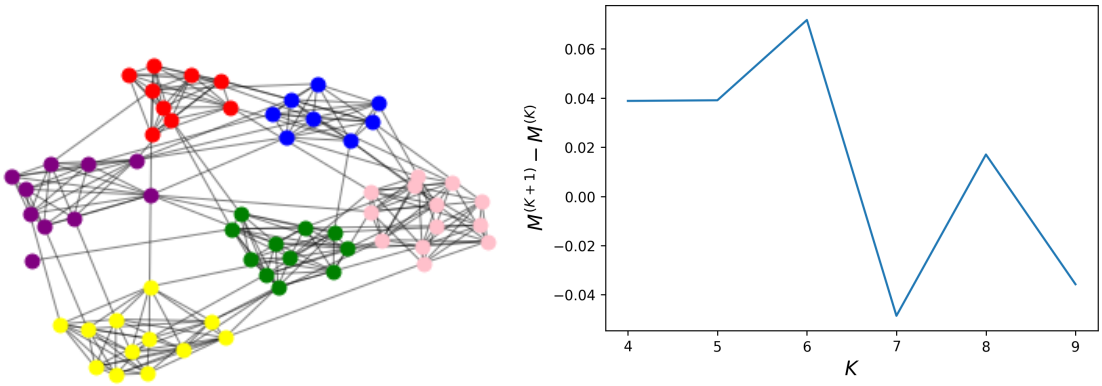


Figure 2.7: $K \mapsto M^{(K+1)} - M^{(K)}$ working with the connectivity matrix Q from (2.1).



(a) Graph.

(b) $K \mapsto M^{(K+1)} - M^{(K)}$

Figure 2.8: We test our model selection method on the *football network* from the Networkx python package.

2.6.3 Application on real data

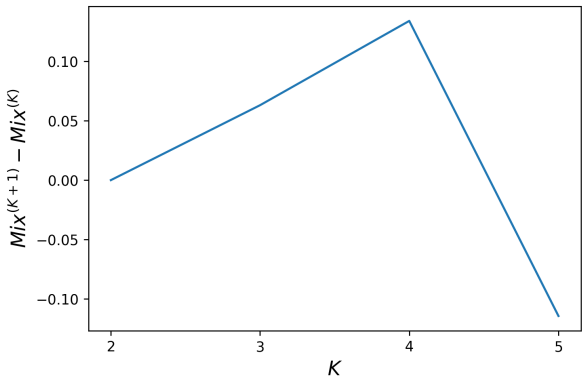
Migratory animals are essential components of the ecosystems that support all life on Earth. By acting as pollinators and seed distributors they contribute to ecosystem structure and function. They provide food for other animals and regulate the number of species in ecosystems. Migratory animals are potentially very effective indicators of environmental changes that affect us all.
In Kölzsch et al. [2018], the authors proposed a periodic Markov model on a spatial migration network to formally describe the process of animal migration on the population level. They built their dataset using the Movebank data repository (see Kruckenberg et al. [2018]) that provides historic of animal movements. We propose to test our approach on this dataset. The data is publicly available here[3] and our experiments can be reproduced with the notebook *experiments.ipynb*.

Description of the dataset. The dataset presents the locations of several white-fronted gooses with timestamps. The animals have been tracked from 2006 to 2010. Each location can be associated with a class using classes defined from Argos User's Manual 2011. We refer to Kölzsch et al. [2018] for details. We focus on one specific white-fronted goose and we keep the list of its chronological locations between 2006 and 2010 for four location classes. Nodes correspond to the entries of the previous sequence of locations of the animal while communities are the classes associated to each location. In our network, we connect two nodes if the distance between the corresponding precise locations (given with latitude and longitude coordinates) is smaller than some specified threshold.

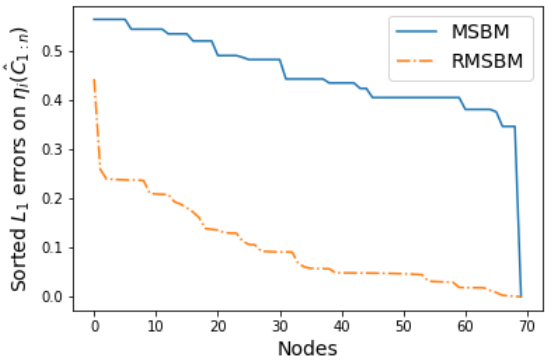
Results. With Figure 2.9.(a), we show that the model selection method of Section 2.6.2 allows to retrieve the correct number of clusters on our dataset. In order to evaluate the performance of our reliable link prediction method, we compute the transition matrix P and the connection matrix Q associated with our network. More precisely, we define $\forall k, l \in [K]$,

$$Q_{k,l} := \begin{cases} \dfrac{1}{|G_k| \times |G_l|} \displaystyle\sum_{i \in G_k} \sum_{j \in G_l} X_{i,j} & \text{if } k \neq l \\ \dfrac{1}{|G_k| \times (|G_k| - 1)} \displaystyle\sum_{i,j \in G_k} X_{i,j} & \text{if } k = l \end{cases} \quad \text{and} \quad P_{k,l} := \frac{n}{n-1} \frac{\sum_{i=1}^{n-1} \mathbb{1}_{(C_i, C_{i+1}) = (k,l)}}{\sum_{i=1}^{n} \mathbb{1}_{C_i = k}},$$

where X is the adjacency matrix, C_i is the community of node $i \in [n]$ and G_k is the set of nodes with label $k \in [K]$. We use these matrices to compute the posterior probabilities $\left(\eta_i(\hat{\mathbf{C}}_{1:n})\right)_{i \in [n]}$ (see Definition 2.6) and we can compare them with the estimations given by the plug-in approach and the reliable approach from Section 2.4. Figure 2.9.(b) shows that the reliable approach allows to significantly improve the estimate of the posterior probabilities.



(a) $K \mapsto M^{(K+1)} - M^{(K)}$.



(b) L^1 errors between $\hat{\eta}_i(\hat{\mathbf{C}}_{1:n})$ (resp. $\hat{\eta}_i^R(\hat{\mathbf{C}}_{1:n})$) and $\eta_i(\hat{\mathbf{C}}_{1:n})$.

Figure 2.9: We test our model on the bird migrations dataset from Kölzsch et al. [2018].

Comments. On simulated data with a small number of clusters, when n gets larger, the clustering algorithm will recover (almost) perfectly the true partition. In that case, it is clear that the reliable version cannot improve drastically the plug in method since this latter (almost) coincides with the Bayes estimator. However, real datasets never fit a particular model and recovering the true partition is really unlikely even for very large graphs. In such cases, our method is of great interest to provide reliable estimations for link prediction despite clustering errors.

[3]https://www.datarepository.movebank.org/handle/10255/move.747

2.7 Proofs

2.7.1 Proof of Theorem 2.2

Lemma 2.12 provides a more complete version of Theorem 2.2 by giving explicitly the constants.

Lemma 2.12. *Let us consider the three positive constants $c,\ c'$ and c'' involved in Theorem 2.17. Assume that $\alpha_n \log(n) \leq 1/L$ and that $n\alpha_n > \max\left(\frac{4Lc''}{\pi_m^2 D^2}\ ,\ \frac{2}{L\pi_m}\right)$. Then it holds*

$$\mathbb{P}\left(\mathrm{err}(\hat{G}, G) > \exp\left(-\frac{c' S^2}{2}\right)\right) \leq \frac{c}{n^2} + 2K \exp\left(-\frac{n\pi_m^2}{2A_1 + 4A_2\pi_m}\right),$$

where $S^2 = \frac{n\alpha_n \pi_m D^2}{L}$ and where A_1 and A_2 are constants that only depend on the Markov chain $(C_i)_{i\geq 1}$ with $A_1 := \frac{1 + (\lambda_+ \vee 0)}{1 - (\lambda_+ \vee 0)}$ and $A_2 := \frac{1}{3}\mathbb{1}_{\lambda_+ \leq 0} + \frac{5}{1-\lambda_+}\mathbb{1}_{\lambda_+ > 0}$. Here $1 - \lambda_+$ is the right L^2 spectral gap of the Markov chain $(C_i)_{i\geq 1}$ (see Definition A.12 in Section A.3).

Remarks.

- The fact that $\pi_m > 0$ is a direct consequence of the positive recurrent property of the Markov chain.
- The second term in the right hand side of the inequality from Lemma 2.12 comes from the concentration of the average number of visits of the Markov chain towards the stationary distribution of the chain. The first term in this inequality corresponds to the bound from Theorem 2.17 when communities have been assigned.
  Recalling that $\|Q\|_\infty$ is upper bounded by $\alpha_n L$, the condition $\alpha_n \log(n) \leq 1/L$ enforces the signal to noise ratio defined by Giraud and Verzelen $s^2 := \Delta^2/(\alpha_n L)$ (see Theorem 2.17) to be larger than $\Delta^2 \times \log(n)$. Another way to interpret this condition is to say that it enforces the expected degree of all nodes of the graph to be smaller than $n/\log(n)$.
- In order to get some intuition on the conditions on n in the previous theorem, keep in mind that asymptotically, the size of the smallest community in the graph will be $n \times \pi_m$.
  - The condition $n > \frac{4Lc''}{\alpha_n \pi_m^2 D^2}$ can be read as $(n \times \pi_m)\alpha_n D^2/L > \frac{4c''}{\pi_m} = 4c'' \times \frac{n}{n\pi_m}$. Asymptotically, $(n \times \pi_m)\alpha_n D^2/L$ provides a lower bound on the signal-to-noise ratio defined in Theorem 2.17. This shows that the condition $n > \frac{4Lc''}{\alpha_n \pi_m^2 D^2}$ is related to the constraint $s^2 \gtrsim n/m$ of Theorem 2.17.
  - The condition $n > \frac{2}{\alpha_n L \pi_m}$ can be read as $\frac{1}{n \times \pi_m} < \alpha_n L/2$. This shows that the condition $n > \frac{2}{\alpha_n L \pi_m}$ is related to the constraint $1/m < \alpha_n L$ from Theorem 2.17.

The proof of Lemma 2.12 is based on the following Lemma which is proved at the end of this subsection.

Lemma 2.13. *We consider c, c' and c'' the three numerical constants involved in Theorem 2.17. Let us consider $0 < t < \pi_m$. Assume that $\alpha_n L \leq 1/\log(n)$. Then for any $\epsilon > 0$ and n large enough such that:*

$$n \times (\pi_m - t) \geq \begin{cases} \frac{L\log(1/\epsilon)}{c'\alpha_n D^2} & (i) \\ \left(\frac{c''nL}{\alpha_n D^2}\right)^{1/2} & (ii) \\ 1/(\alpha_n L) & (iii) \end{cases}$$

it holds

$$\mathbb{P}\left(\mathrm{err}(\hat{\mathrm{G}}, \mathrm{G}) > \epsilon\right) \leq \frac{c}{n^2} + 2K \exp\left(-\frac{nt^2}{2(A_1/4 + A_2 t)}\right),$$

where A_1 and A_2 are constants defined in Theorem 2.2.

Note that the only constraint on ϵ is given by the condition (i) which is equivalent to

$$\epsilon \geq \exp\left(-\frac{c' D^2 n\alpha_n(\pi_m - t)}{L}\right).$$

In order to get the tighter result possible, we want to choose $\epsilon = \exp\left(-\frac{c' D^2 n\alpha_n(\pi_m - t)}{L}\right)$ which leads to

$$t = \pi_m - \frac{L\log(1/\epsilon)}{c' D^2 n\alpha_n}.$$

The condition $t > 0$ is then equivalent to

$$\pi_m > \frac{L\log(1/\epsilon)}{c' D^2 n\alpha_n} \Leftrightarrow \exp(-\pi_m n\alpha_n c' D^2/L) < \epsilon.$$

The condition (ii) is equivalent to

$$n(\pi_m - t) = \frac{L\log(1/\epsilon)}{c'\alpha_n D^2} \geq \left(\frac{c'' nL}{\alpha_n D^2}\right)^{1/2} \Leftrightarrow \exp\left(-c'\sqrt{\frac{D^2 c'' n\alpha_n}{L}}\right) \geq \epsilon.$$

The condition (iii) is equivalent to

$$n(\pi_m - t) = \frac{L\log(1/\epsilon)}{c'\alpha_n D^2} \geq (1/\alpha_n L) \Leftrightarrow \exp\left(-\frac{c' D^2}{L^2}\right) \geq \epsilon.$$

One can easily prove that for $n\alpha_n > \max\left(\frac{4Lc''}{\pi_m^2 D^2}, \frac{2}{L\pi_m}\right)$, $\epsilon := \exp\left(-\frac{\pi_m n\alpha_n c' D^2}{2L}\right)$ satisfies the three conditions above. This gives Lemma 2.12 from Lemma 2.13.

Proof of Lemma 2.13. Using Theorem 1.2 from Jiang et al. [2018], we get that

$$\forall c \in [K],\ \forall t > 0,\ \ \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i = c} - \pi(c)\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(A_1\sigma_c^2 + A_2 t)}\right) \tag{2.6}$$

where $A_1 = \frac{1 + (\lambda_+ \vee 0)}{1 - (\lambda_+ \vee 0)}$, $A_2 = \frac{1}{3}\mathbb{1}_{\lambda_+ \leq 0} + \frac{5}{1 - \lambda_+}\mathbb{1}_{\lambda_+ > 0}$ and $\sigma_c^2 = \pi(c)(1 - \pi(c))$.
We deduce that for all $t > 0$,

$$\mathbb{P}\left(\bigcup_c \left\{\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i = c} - \pi(c)\right| \geq t\right\}\right) \leq 2K\exp\left(-\frac{nt^2}{2(A_1\sigma^2 + A_2 t)}\right),$$

where $\sigma^2 := \max_c \sigma_c^2 \ \ (\leq 1/4)$. We define $\Omega^c := \bigcup_c \left\{\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i = c} - \pi(c)\right| \geq t\right\}$ and we recall $\pi_m = \min_c \pi(c)$ and $D^2 = \min_{j \neq k} \sum_l ((Q_0)_{k,l} - (Q_0)_{j,l})^2$.
Suppose that $0 < t < \pi_m$ and that n is large enough to satisfy (i), (ii) and (iii). Then it holds

$$\begin{aligned}
& \mathbb{P}\left(\text{err}(\hat{G}, G) > \epsilon\right) \\
= \ & \mathbb{P}\left(\{\text{err}(\hat{G}, G) > \epsilon\} \cap \Omega\right) + \mathbb{P}\left(\{\text{err}(\hat{G}, G) > \epsilon\} \cap \Omega^c\right) \\
\leq \ & \mathbb{P}\left(\{\text{err}(\hat{G}, G) > \epsilon\} \cap \Omega\right) + 2K\exp\left(-\frac{nt^2}{2(A_1\sigma^2 + A_2 t)}\right) \\
= \ & \mathbb{P}\left(\text{err}(\hat{G}, G) > \epsilon \mid \Omega\right) \times \mathbb{P}\left(\Omega\right) + 2K\exp\left(-\frac{nt^2}{2(A_1\sigma^2 + A_2 t)}\right).
\end{aligned} \tag{2.7}$$

We denote by M the random variable that gives the size of the smallest cluster: $M := \min_{k \in [K]} \ m_k$.

Condition (i) is equivalent to

$$\epsilon \geq \exp\left(-c' \frac{n\alpha_n(\pi_m - t)D^2}{L}\right).$$

Since on the event Ω we have $n(\pi_m - t) \leq M$, we get that on Ω it holds

$$\epsilon \geq \exp\left(-c' \frac{M\alpha_n D^2}{L}\right) \geq \exp\left(-c' s^2\right), \tag{2.8}$$

where $s^2 = \Delta^2/(\alpha_n L)$ with $\Delta^2 = \min_{k\neq j} \Delta_{k,j}^2$ and $\Delta_{k,j}^2 = \sum_l m_l(Q_{k,l} - Q_{j,l})^2$. The last inequality comes from the fact that $\Delta^2 \geq M\alpha_n^2 D^2$. Using (2.7) we get that

$$\begin{aligned}\mathbb{P}\left(\mathrm{err}(\hat{G}, G) > \epsilon\right) &\leq \mathbb{P}\left(\mathrm{err}(\hat{G}, G) > \epsilon \mid \Omega\right) + 2K\exp\left(-\frac{nt^2}{2(A_1\sigma^2 + A_2 t)}\right) \\ &\leq \mathbb{P}\left(\mathrm{err}(\hat{G}, G) > e^{-c's^2} \mid \Omega\right) + 2K\exp\left(-\frac{nt^2}{2(A_1\sigma^2 + A_2 t)}\right).\end{aligned}$$

We note that on Ω :

- Condition (ii) gives
$$M^2 \geq \frac{c''nL}{\alpha_n D^2} \Leftrightarrow \frac{M\alpha_n D^2}{L} \geq c''n/M,$$
which implies that $s^2 = \frac{\Delta^2}{\alpha_n L} \geq c''n/M$ since $\Delta^2 \geq M\alpha_n^2 D^2$.
- Condition (iii) gives
$$\frac{1}{M} \leq \alpha_n L.$$

Applying Theorem 2.17 from Verzelen and Giraud, we get that

$$\mathbb{P}\left(\mathrm{err}(\hat{G}, G) > e^{-c's^2} | \Omega\right) \leq \frac{c}{n^2}.$$

Finally we obtain using Eq.(2.8) that

$$\mathbb{P}\left(\mathrm{err}(\hat{G}, G) > \epsilon\right) \leq \frac{c}{n^2} + 2K\exp\left(-\frac{nt^2}{2(A_1\sigma^2 + A_2 t)}\right).$$

□

2.7.2 Proof of Theorem 2.3

We start by proving Lemma 2.14 which enriches the statement of Theorem 2.3 by giving explicitly the constants.

Lemma 2.14. *We consider c, c' and c'' the three numerical constants involved in Theorem 2.17. Assume that $\alpha_n \log(n) \leq \frac{1}{L}$ and that $n\alpha_n > \max\left(\frac{4Lc''}{\pi_m^2 D^2}, \frac{2}{L\pi_m}, \frac{2L\log(n)}{\pi_m c' D^2}\right)$. Then for all $0 < t < \pi_m - \frac{1}{n}$, it holds*

$$\begin{aligned}&\mathbb{P}\left(\|\hat{Q} - Q\|_\infty > t\right) \\ \leq\ & K(K+1)\exp\left(-\frac{(n\pi_m - nt - 1)^2 t^2}{\frac{1}{2} + \frac{2}{3}t}\right) + \frac{c}{n^2} + 2K\exp\left(-\frac{nt^2}{2(A_1/4 + A_2 t)}\right).\end{aligned}$$

Proof of Lemma 2.14.

- Preliminary 1

Using the standard Bernstein's inequality for independent random variables, we get that for all $k, l \in [K]^2$ with $k \neq l$, and for all $t > 0$, it holds

$$\mathbb{P}\left(\left|\frac{1}{|G_k| \times |G_l|}\sum_{i\in G_k}\sum_{j\in G_l} X_{i,j} - Q_{k,l}\right| \geq t\right) \leq 2\exp\left(-\frac{|G_k| \times |G_l| t^2}{2(Q_{k,l}(1-Q_{k,l}) + t/3)}\right)$$

and for all $k \in [K], t > 0$, it holds

$$\mathbb{P}\left(\left|\frac{1}{|G_k| \times (|G_k|-1)}\sum_{i,j\in G_k\ i\neq j} X_{i,j} - Q_{k,k}\right| \geq t\right) \leq 2\exp\left(-\frac{|G_k| \times (|G_k|-1) t^2}{2(Q_{k,k}(1-Q_{k,k}) + t/3)}\right).$$

- Preliminary 2

  We define the event $N := \left\{\text{err}(\hat{G}, G) < \exp\left(-\frac{\pi_m n \alpha_n c' D^2}{2L}\right)\right\}$. Note that on N, the partition of the clusters is correctly recovered thanks to the condition $n\alpha_n > \frac{2L\log(n)}{\pi_m c' D^2}$.

- Preliminary 3

  Using Theorem 1.2 from Jiang et al. [2018], we get that

  $$\forall c \in [K],\ \forall t > 0,\ \ \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i=c} - \pi(c)\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(A_1/4 + A_2 t)}\right).$$

  We deduce that for all $t > 0$,

  $$\mathbb{P}\left(\bigcup_c \left\{\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i=c} - \pi(c)\right| \geq t\right\}\right) \leq 2K\exp\left(-\frac{nt^2}{2(A_1/4 + A_2 t)}\right).$$

  We define $\Omega^c := \bigcup_{c\in[K]} \left\{\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i=c} - \pi(c)\right| \geq t\right\}$.

Considering $0 < t < \pi_m - \frac{1}{n}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\|\hat{Q} - Q\|_\infty > t\right) \\
\leq\ &\mathbb{P}\left(\bigcup_{k,l\in[K]^2,\ k\leq l} \{|\hat{Q}_{k,l} - Q_{k,l}| > t\} \mid \Omega\right) + \mathbb{P}(\Omega^c) \\
\leq\ &\mathbb{P}\left(\bigcup_{k,l\in[K]^2,\ k\leq l} \{|\hat{Q}_{k,l} - Q_{k,l}| > t\} \mid N, \Omega\right) + \mathbb{P}(N^c \mid \Omega) + \mathbb{P}(\Omega^c) \\
&\text{and using preliminary 3,} \\
\leq\ &\mathbb{P}\left(\bigcup_{k,l\in[K]^2,\ k\leq l} \{|\hat{Q}_{k,l} - Q_{k,l}| > t\} \mid N, \Omega\right) + \mathbb{P}(N^c \mid \Omega) \\
&+ 2K\exp\left(-\frac{nt^2}{2(A_1/4 + A_2 t)}\right) \\
\leq\ &\mathbb{P}\left(\bigcup_{k,l\in[K]^2,\ k\leq l} \{|\hat{Q}_{k,l} - Q_{k,l}| > t\} \mid N, \Omega\right) + \frac{c}{n^2} + 2K\exp\left(-\frac{nt^2}{2(A_1/4 + A_2 t)}\right) \\
&\text{where we used that } \mathbb{P}(N^c \mid \Omega) \leq \frac{c}{n^2} \text{ (shown in the proof of Theorem 2.2),}
\end{aligned}$$

$$\leq \quad 2 \sum_{1\leq k\leq l\leq K} \exp\left(-\frac{n(\pi(k)-t)\times(n\pi(l)-nt-1)t^2}{2(Q_{k,l}(1-Q_{k,l})+t/3)}\right)+\frac{c}{n^2}+$$
$$2K\exp\left(-\frac{nt^2}{2(A_1/4+A_2t)}\right),$$

where the last inequality is a direct consequence of the three preliminaries. □

Proof of Theorem 2.3. Let us consider $\gamma>0$ and let us define $t=\frac{\gamma}{\sqrt{n}}$. Considering that

$$\frac{n\alpha_n}{\log(n)}\geq a \quad \text{with} \quad a:=\frac{4Lc''}{c'\pi_m^2D^2}\vee\frac{2L}{c'\pi_mD^2}\vee\frac{2}{L\pi_m},$$

we ensure that $n\alpha_n$ satisfies the conditions of Lemma 2.14.
Now let us look into the condition $t=\frac{\gamma}{\sqrt{n}}<\pi_m-\frac{1}{n}$ of Lemma 2.14. We will ask t to satisfy the stronger condition

$$t=\frac{\gamma}{\sqrt{n}}<\frac{\pi_m}{2}-\frac{1}{n} \quad \Leftrightarrow \quad 0<\frac{\pi_m}{2}n-\gamma\sqrt{n}-1. \tag{2.9}$$

Studying the polynomial function $f:x\mapsto\frac{\pi_m}{2}x^2-\gamma x-1$, one can find that the zeros of f are

$$x_1:=\frac{\gamma-\sqrt{\gamma^2+2\pi_m}}{\pi_m} \quad \text{and} \quad x_2:=\frac{\gamma+\sqrt{\gamma^2+2\pi_m}}{\pi_m}\leq\frac{2\gamma+\sqrt{2\pi_m}}{\pi_m}.$$

We deduce that considering that

$$n>4\left(\frac{\gamma+1}{\pi_m}\right)^2, \tag{2.10}$$

which implies that $\sqrt{n}>\frac{2\gamma+\sqrt{2\pi_m}}{\pi_m}$, we guarantee that $\gamma/\sqrt{n}<\pi_m-1/n$. Applying Lemma 2.14, we get that with probability at least

$$1-\left[(K^2+K)\exp\left(\frac{-(n\pi_m-\gamma\sqrt{n}-1)^2\frac{\gamma^2}{n}}{\frac{1}{2}+\frac{2}{3}\frac{\gamma}{\sqrt{n}}}\right)+\frac{c}{n^2}+2K\exp\left(\frac{-\gamma^2}{2(A_1/4+A_2\frac{\gamma}{\sqrt{n}})}\right)\right],$$

it holds $\|\hat{Q}-Q\|_\infty\leq\gamma/\sqrt{n}$.
Thanks to Eqs.(2.10) and (2.9), we have $(n\pi_m-\gamma\sqrt{n}-1)^2=n^2(\pi_m-\gamma/\sqrt{n}-1/n)^2\geq n^2\pi_m^2/4$ and $\gamma/\sqrt{n}\leq\pi_m/2$. We deduce that defining

$$b:=c+(2K(K+1)) \quad \text{and} \quad b':=\frac{1}{2(A_1/4+A_2\pi_m)}\wedge\frac{\pi_m^2}{2+\frac{4}{3}\pi_m},$$

it holds with probability at least $1-b(1/n^2\vee\exp(-b'\gamma^2))$

$$\|\hat{Q}-Q\|_\infty\leq\frac{\gamma}{\sqrt{n}}.$$

□

2.7.3 Proof of Theorem 2.4

Lemma 2.15 provides a more complete version of Theorem 2.4 by giving explicitly the constants.

Lemma 2.15. *We consider c, c' and c'' the three numerical constants involved in Theorem 2.17. Assume that $\alpha_n\log(n)\leq 1/L$ and that $n\alpha_n>\max\left(\frac{4Lc''}{\pi_m^2D^2}\,,\,\frac{2}{L\pi_m}\,,\,\frac{2L\log(n)}{\pi_mc'D^2}\right)$. Then for all $t>0$, it holds*

$$\mathbb{P}\left(\|\hat{\pi}-\pi\|_\infty>t\right)\leq 2K\exp\left(-\frac{nt^2}{2(A_1/4+A_2t)}\right)+\frac{c}{n^2}+2K\exp\left(-\frac{n\pi_m^2}{2A_1+4A_2\pi_m}\right).$$

Proof of Lemma 2.15. Using Theorem 1.2 from Jiang et al. [2018], we get that

$$\forall c \in [K],\ \forall t > 0,\ \ \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i=c} - \pi(c)\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(A_1\sigma_c^2 + A_2 t)}\right)$$

where $A_1 = \dfrac{1+(\lambda_+ \vee 0)}{1-(\lambda_+ \vee 0)}$, $A_2 = \dfrac{1}{3}\mathbb{1}_{\lambda_+ \leq 0} + \dfrac{5}{1-\lambda_+}\mathbb{1}_{\lambda_+ > 0}$ and $\sigma_c^2 = \pi(c)(1-\pi(c)) \leq 1/4$.

We define the event $N := \left\{\mathrm{err}(\hat{G}, G) < \exp\left(-\frac{\pi_m n \alpha_n c' D^2}{2L}\right)\right\}$. Note that on N, the partition of the clusters is correctly recovered thanks to the condition $n\alpha_n > \frac{2L\log(n)}{\pi_m c' D^2}$. Then,

$$\begin{aligned}
& \mathbb{P}\left(\bigcup_{k\in[K]} \{|\hat{\pi}(k) - \pi(k)| > t\}\right) \\
\leq\ & \mathbb{P}\left(\bigcup_{k\in[K]} \{|\hat{\pi}(k) - \pi(k)| > t\} \mid N\right) + \mathbb{P}(N^c) \\
=\ & \mathbb{P}\left(\bigcup_{k\in[K]} \{|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_i=k} - \pi(k)| > t\} \mid N\right) + \mathbb{P}(N^c) \\
\leq\ & 2K\exp\left(-\frac{nt^2}{2(A_1/4 + A_2 t)}\right) + \frac{c}{n^2} + 2K\exp\left(-\frac{n\pi_m^2}{2A_1 + 4A_2\pi_m}\right),
\end{aligned}$$

where we apply Lemma 2.12 in the last inequality. □

2.7.4 Proof of Theorem 2.5

We will prove a more accurate result with Lemma 2.16.

Lemma 2.16. *Let us consider* $\gamma > \frac{5K}{2\pi_m^2}$.

Assume that $\alpha_n \log(n) \leq 1/L$, *that* $n\alpha_n > \max\left(\dfrac{4Lc''}{\pi_m^2 D^2}, \dfrac{4}{L\pi_m}, \dfrac{2L\log(n)}{\pi_m c' D^2}\right)$ *and that* $\sqrt{n} > \frac{2}{\pi_m}(1 + \pi_m^2\gamma/5)$. *Then it holds*

$$\begin{aligned}
& \mathbb{P}\left(\|\hat{P} - P\|_\infty \geq \frac{\gamma}{\sqrt{n}}\right) \\
\leq\ & 2K^2\exp\left(-\frac{\left(\frac{\pi_m^2\gamma}{5K} - \frac{1}{2}\right)^2}{2(B_1/4 + B_2\frac{\pi_m^2\gamma}{5K\sqrt{n}})}\right) + \frac{c}{n^2} + 2K\exp\left(-\frac{n\pi_m^2}{8A_1\sigma^2 + 4A_2\pi_m}\right),
\end{aligned}$$

where B_1 *and* B_2 *depend only on the Markov chain and are defined by* $B_1 := \dfrac{1+(\xi_+ \vee 0)}{1-(\xi_+ \vee 0)}$ *and* $B_2 := \dfrac{1}{3}\mathbb{1}_{\xi_+ \leq 0} + \dfrac{5}{1-\xi_+}\mathbb{1}_{\xi_+ > 0}$. *Here* $1-\xi_+$ *is the right* L^2 *spectral gap of the Markov chain* $(Y_i)_{i\geq 1}$ *(see Definition A.12 in Section A.3).*

Remarks.

- The first term in the right hand side of the inequality in Lemma 2.16 is due to the concentration of the average number of visits of the chain $(Y_i)_{i\geq 1}$ (defined in Section 2.2.3.3 of this chapter) towards its stationary distribution. The two last terms of the inequality correspond to the bound guaranteeing the recovery of the true partitions with a direct application of Theorem 2.2.
- The condition $n\alpha_n > \frac{2L\log(n)}{\pi_m c' D^2}$ ensures that $\exp\left(-\frac{\pi_m n\alpha_n c' D^2}{2L}\right) < \frac{1}{n}$. Theorem 2.2 will then guarantee that we recover perfectly the partition of the communities.

- Expecting the accuracy $\gamma/\sqrt{n}$, the condition $\sqrt{n} > \frac{2}{\pi_m}(1+\pi_m^2\gamma/5)$ ensures that the Markov chain $(C_i)_{i\geq 1}$ has visited enough each state $k \in [K]$ to guarantee the convergence of the average number of visits toward the stationary distribution.

Proof of Lemma 2.16.

I. Concentration of the average number of visits for $(Y_i)_{i\geq 1}$.

We recall that $(Y_i)_{i\geq 1}$ is a Markov Chain on $[K]^2$ defined by : $Y_i = (C_i, C_{i+1})$.
Then using again Theorem 1.2 from Jiang et al. [2018], we get that $\forall t > 0,\ \forall k, l \in [K]^2$,

$$\mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)} - \pi(k)P_{k,l}\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(B_1/4+B_2t)}\right),$$

II. First step toward the Theorem.

We define the event $N := \left\{\text{err}(\hat{G}, G) < \exp\left(-\frac{\pi_m n\alpha_n c' D^2}{2L}\right)\right\}$. Note that on N, the partition of the clusters is correctly recovered thanks to the condition $n\alpha_n > \frac{2L\log(n)}{\pi_m c' D^2}$. Let $\gamma > \frac{5K}{2\pi_m^2}$ and let us define

$$r = \frac{\zeta}{\sqrt{n}} \text{ with } \zeta = \frac{\pi_m^2\gamma}{5K} - \frac{1}{2} > 0,$$
$$\text{and} \quad \Gamma = \bigcap_{k,l}\left\{\left|\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)} - \pi(k)P_{k,l}\right| < r\right\}.$$

Then,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{k,l}\left\{\left|\hat{P}_{k,l} - P_{k,l}\right| \geq \frac{\gamma}{\sqrt{n}}\right\}\right) \\
\leq\ &\underbrace{\mathbb{P}\left(\bigcup_{k,l}\left\{\left|\hat{P}_{k,l} - P_{k,l}\right| \geq \frac{\gamma}{\sqrt{n}}\right\} | N, \Gamma\right)}_{(*)}\mathbb{P}(N)\mathbb{P}(\Gamma|N) + \mathbb{P}(\Gamma^c) + \mathbb{P}(N^c).
\end{aligned}$$

Note that the condition $\sqrt{n} > \frac{2}{\pi_m}(1+\pi_m^2\gamma/5)$ of Lemma 2.16 implies that

$$\sqrt{n} > \frac{2}{\pi_m}(1+K\zeta). \tag{2.11}$$

III. We prove that $(*)$ is zero.

In this third step of the proof, we are going to show that conditionally on the event $N \cap \Gamma$, the infinite norm between our estimate of the transition matrix $\hat{P}$ and P is smaller than $\gamma/\sqrt{n}$.

1 We split $(*)$ in two terms.

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{k,l}\left\{\left|\hat{P}_{k,l} - P_{k,l}\right| \geq \frac{\gamma}{\sqrt{n}}\right\} | N, \Gamma\right) \\
=\ &\mathbb{P}\left(\bigcup_{k,l}\left\{\left|\hat{P}_{k,l} - \frac{\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)}}{(n-1)\pi(k)} + \frac{\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)}}{(n-1)\pi(k)} - P_{k,l}\right| \geq \frac{\gamma}{\sqrt{n}}\right\} | N, \Gamma\right)
\end{aligned}$$

2 We show that on Γ: $\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_i=k} - \pi(k)\right| \leq \frac{1}{n} + Kr$. Here we show that a concentration of the average number of visits for $(Y_i)_{i\geq 1}$ gives for free a concentration result of the average number of visits

for $(C_i)_{i\geq 1}$.

Note that on the event Γ :

$$\begin{aligned}
\bullet\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_i=k} &= \frac{1}{n}\sum_{l=1}^{K}\sum_{i=1}^{n-1}\mathbb{1}_{C_i=k,C_{i+1}=l}\\
&= \frac{n-1}{n}\sum_{l=1}^{K}\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{1}_{C_i=k,C_{i+1}=l}\\
&\geq \frac{n-1}{n}\sum_{l=1}^{K}(\pi(k)P_{k,l}-r)\\
&= \frac{n-1}{n}(\pi(k)-Kr).
\end{aligned}$$

Hence $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_i=k}-\pi(k)\geq -\frac{\pi(k)}{n}-\frac{n-1}{n}Kr\geq -\left(\frac{1}{n}+Kr\right)$.

$$\begin{aligned}
\bullet\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_i=k} &\leq \frac{1}{n}\sum_{l=1}^{K}\sum_{i=1}^{n-1}\mathbb{1}_{C_i=k,C_{i+1}=l}+\frac{1}{n}\\
&= \frac{n-1}{n}\sum_{l=1}^{K}\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{1}_{C_i=k,C_{i+1}=l}+\frac{1}{n}\\
&\leq \frac{n-1}{n}\sum_{l=1}^{K}(\pi(k)P_{k,l}+r)+\frac{1}{n}\\
&\leq \pi(k)+Kr+\frac{1}{n}.
\end{aligned}$$

Hence $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_i=k}-\pi(k)\leq \frac{1}{n}+Kr$.

We deduce then that on Γ, $\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_i=k}-\pi(k)\right|\leq \frac{1}{n}+Kr$.

3 **We show that the first term from** **1** **is zero.** In the following, we show that the definition of ζ with the condition $\sqrt{n}>\frac{2}{\pi_m}(1+\pi_m^2\gamma/5)$ implies that the first term in 1 is zero.

$$\begin{aligned}
& \mathbb{P}\left(\left|\hat{P}_{k,l}-\frac{1}{n-1}\frac{\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)}}{\pi(k)}\right|\geq\frac{\gamma}{2\sqrt{n}}\mid N,\Gamma\right)\\
= & \mathbb{P}\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)}\left|\frac{n}{\sum_{i=1}^{n}\mathbb{1}_{C_i=k}}-\frac{1}{\pi(k)}\right|\geq\frac{\gamma}{2\sqrt{n}}\mid N,\Gamma\right)\\
\leq & \mathbb{P}\left((r+\pi(k)P_{k,l})\left|\frac{n}{\sum_{i=1}^{n}\mathbb{1}_{C_i=k}}-\frac{1}{\pi(k)}\right|\geq\frac{\gamma}{2\sqrt{n}}\mid N,\Gamma\right)\quad\text{(by definition of }\Gamma)\\
= & \mathbb{P}\left(\left|\frac{n\pi(k)-\sum_{i=1}^{n}\mathbb{1}_{C_i=k}}{\pi(k)\sum_{i=1}^{n}\mathbb{1}_{C_i=k}}\right|\geq\frac{1}{2r+2\pi(k)P_{k,l}}\cdot\frac{\gamma}{\sqrt{n}}\mid N,\Gamma\right)\text{ and using }\boxed{2},\\
\leq & \mathbb{P}\left(\frac{\frac{1}{n}+Kr}{\pi(k)(\pi(k)-\frac{1}{n}-Kr)}\geq\frac{1}{2r+2\pi(k)P_{k,l}}\cdot\frac{\gamma}{\sqrt{n}}\mid N,\Gamma\right)\\
\leq & \mathbb{P}\left(\frac{\frac{1}{n}+Kr}{\pi_m(\pi_m-\frac{1}{n}-Kr)}\geq\frac{1}{2r+2}\cdot\frac{\gamma}{\sqrt{n}}\mid N,\Gamma\right)\text{ and since }r=\frac{\zeta}{\sqrt{n}},
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(\frac{\frac{1}{n}+K\frac{\zeta}{\sqrt{n}}}{\pi_m(\pi_m-\frac{1}{n}-K\frac{\zeta}{\sqrt{n}})} \geq \frac{1}{2\frac{\zeta}{\sqrt{n}}+2}\cdot\frac{\gamma}{\sqrt{n}} \mid N,\Gamma\right)\\
&= \mathbb{P}\left(\frac{(\frac{1}{n}+K\frac{\zeta}{\sqrt{n}})(2\zeta+2\sqrt{n})}{\pi_m(\pi_m-\frac{1}{n}-K\frac{\zeta}{\sqrt{n}})} \geq \gamma \mid N,\Gamma\right)\\
&\leq \mathbb{P}\left(\frac{(\frac{1}{n}+K\frac{\zeta}{\sqrt{n}})(2\zeta+2\sqrt{n})}{\pi_m(\pi_m-\frac{1}{\sqrt{n}}(1+K\zeta))} \geq \gamma \mid N,\Gamma\right). \qquad (2.12)
\end{aligned}
$$

Since from (2.11), $\sqrt{n} \geq \frac{2}{\pi_m}(1+K\zeta)$, we have

$$\frac{\pi_m}{2} \leq \pi_m - \frac{1}{\sqrt{n}}(1+K\zeta),$$

which leads to

$$\mathbb{P}\left(\frac{(\frac{1}{n}+K\frac{\zeta}{\sqrt{n}})(2\zeta+2\sqrt{n})}{\pi_m(\pi_m-\frac{1}{\sqrt{n}}(1+K\zeta))} \geq \gamma \mid N,\Gamma\right) \leq \mathbb{P}\left(\frac{2(\frac{1}{\sqrt{n}}+K\zeta)(\frac{\zeta}{\sqrt{n}}+1)}{\pi_m^2/2} \geq \gamma \mid N,\Gamma\right).$$

Moreover, since from (2.11) and the fact that $\pi_m \in (0,1)$, $\sqrt{n} \geq \frac{2}{\pi_m}(1+K\zeta) > 2K\zeta$, it holds

$$\frac{\zeta}{\sqrt{n}} < \frac{1}{2K} < \frac{1}{4}.$$

Coming back to (2.12), we finally get

$$
\begin{aligned}
&\mathbb{P}\left(\left|\hat{P}_{k,l} - \frac{1}{n-1}\frac{\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)}}{\pi(k)}\right| \geq \frac{\gamma}{2\sqrt{n}} \mid N,\Gamma\right)\\
\leq\ &\mathbb{P}\left(\frac{2(\frac{1}{\sqrt{n}}+K\zeta)(\frac{\zeta}{\sqrt{n}}+1)}{\pi_m^2/2} \geq \gamma \mid N,\Gamma\right)\\
\leq\ &\mathbb{P}\left(\frac{5(\frac{1}{\sqrt{n}}+K\zeta)}{\pi_m^2} \geq \gamma \mid N,\Gamma\right)\\
=\ &0.
\end{aligned}
$$

The last equality is due to the definition of ζ. Indeed,

$$\zeta = \frac{\gamma\pi_m^2}{5K} - \frac{1}{2} < \frac{\gamma\pi_m^2}{5K} - \frac{1}{K\sqrt{n}} \quad \text{leading to} \quad \frac{5(\frac{1}{\sqrt{n}}+K\zeta)}{\pi_m^2} < \gamma.$$

4 We show that the second term from 1 is zero.

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n-1}\frac{\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)}}{\pi(k)} - P_{k,l}\right| \geq \frac{\gamma}{2\sqrt{n}} \mid N,\Gamma\right)\\
=\ &\mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)} - \pi(k)P_{k,l}\right| \geq \pi(k)\frac{\gamma}{2\sqrt{n}} \mid N,\Gamma\right)\\
\leq\ &\mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{1}_{Y_i=(k,l)} - \pi(k)P_{k,l}\right| \geq \pi_m\frac{\gamma}{2\sqrt{n}} \mid N,\Gamma\right)\\
=\ &0,
\end{aligned}
$$

where the last equality comes from the definition of $r = \zeta/\sqrt{n}$ and the definition of Γ because

$$\zeta = \frac{\pi_m^2\gamma}{5K} - \frac{1}{2} < \frac{\pi_m^2\gamma}{5K} \leq \frac{\pi_m\gamma}{2}.$$

IV. Conclusion.

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{k,l}\left\{\left|\hat{P}_{k,l} - P_{k,l}\right| \geq \frac{\gamma}{\sqrt{n}}\right\}\right) \\
\leq\ &\mathbb{P}\left(\bigcup_{k,l}\left\{\left|\hat{P}_{k,l} - P_{k,l}\right| \geq \frac{\gamma}{\sqrt{n}}\right\} \mid N, \Gamma\right)\mathbb{P}(N)\mathbb{P}(\Gamma \mid N) + \mathbb{P}(\Gamma^c) + \mathbb{P}(N^c) \\
=\ &\mathbb{P}(\Gamma^c) + \mathbb{P}(N^c) \\
\leq\ &2K^2\exp\left(-\frac{nr^2}{2(B_1/4 + B_2 r)}\right) + \frac{c}{n^2} + 2K\exp\left(-\frac{n\pi_m^2}{8A_1\sigma^2 + 4A_2\pi_m}\right) \\
\leq\ &2K^2\exp\left(-\frac{\left(\frac{\pi_m^2\gamma}{5K} - \frac{1}{2}\right)^2}{2(B_1/4 + B_2\frac{\frac{\pi_m^2\gamma}{5K} - \frac{1}{2}}{\sqrt{n}})}\right) + \frac{c}{n^2} + 2K\exp\left(-\frac{n\pi_m^2}{8A_1\sigma^2 + 4A_2\pi_m}\right),
\end{aligned}$$

where we apply Lemma 2.12 in the last inequality. □

2.7.5 Proof of Proposition 2.11

Let us consider $\gamma > \frac{5K}{2\pi_m^2}$. We assume that the conditions (2.4) of Proposition 2.11 are satisfied and we deduce from Theorems 2.5 and 2.3 that there exists three constants $a, b, b' > 0$ such that with probability at least $1 - \frac{1}{2}b\left[1/n^2 \vee \exp\left(-b'(\gamma - \frac{5K}{2\pi_m^2})^2\right)\right]$, it holds

$$\|\hat{P} - P\|_\infty \vee \|\hat{Q} - Q\|_\infty \leq \frac{\gamma}{\sqrt{n}}.$$

For any $i \in [n]$ we have

$$\begin{aligned}
&|\eta_i(\mathbf{C}_{1:n}) - \hat{\eta}_i(\mathbf{C}_{1:n})| \\
&= \left|\sum_{k\in[K]} P_{C_i,k}Q_{C_i,k} - \sum_{k\in[K]} \hat{P}_{C_i,k}\hat{Q}_{C_i,k}\right| \\
&\leq \sum_{k\in[K]}\left|P_{C_i,k}Q_{C_i,k} - \hat{P}_{C_i,k}Q_{C_i,k} + \hat{P}_{C_i,k}Q_{C_i,k} - \hat{P}_{C_i,k}\hat{Q}_{C_i,k}\right| \\
&\leq \sum_{k\in[K]}\left|P_{C_i,k} - \hat{P}_{C_i,k}\right| \times |Q_{C_i,k}| + \sum_{k\in[K]}\left|Q_{C_i,k} - \hat{Q}_{C_i,k}\right| \times \left|\hat{P}_{C_i,k}\right| \\
&\leq \|\hat{P} - P\|_\infty \sum_{k\in[K]} Q_{C_i,k} + \|Q - \hat{Q}\|_\infty \sum_{k\in[K]} \hat{P}_{C_i,k} \\
&\leq \|\hat{P} - P\|_\infty K\alpha_n L + \|Q - \hat{Q}\|_\infty,
\end{aligned}$$

where we used that $\|Q\|_\infty = \alpha_n\|Q_0\|_\infty \leq \alpha_n L$ and the fact that $\hat{P}$ is a stochastic matrix.
We deduce that for any $i \in [n]$, it holds with probability at least $1 - b\left[1/n^2 \vee \exp\left(-b'(\gamma - \frac{5K}{2\pi_m^2})^2\right)\right]$,

$$|\eta_i(\mathbf{C}_{1:n}) - \hat{\eta}_i(\mathbf{C}_{1:n})| \leq \frac{\gamma}{\sqrt{n}}\left(\alpha_n KL + 1\right).$$

Using a union bound concludes the proof.

2.8 Partial recovery bound for SBMs with a SDP method

Partial recovery bound in SBMs with fixed assignment of the communities. In Giraud and Verzelen [2019], the authors introduce a relaxed version of the K-means algorithms on the columns of the adjacency matrix. One specificity of their algorithm is the fact that they are working with the square of the adjacency matrix. This choice allows them to tackle problems outside of the assortative setting and with a wide set of possible connectivity matrices Q contrary to previous works.
Theorem 2.17 presents the result of Verzelen and Giraud in the SBM framework with a connectivity matrix $Q = \alpha_n Q_0$.

Theorem 2.17. *[cf. Giraud and Verzelen, 2019, Theorem 2]*
Assume that $\|Q_0\|_\infty \leq L$. The size of the community $k \in [K]$ will be denoted m_k. The size of the smallest community will be denoted m. We define the signal-to-noise ratio $s^2 = \Delta^2/(\alpha_n L)$, where $\Delta^2 = \min_{k\neq j} \Delta_{k,j}^2$ with $\Delta_{k,j}^2 = \sum_l m_l(Q_{k,l} - Q_{j,l})^2 = \alpha_n^2 \sum_l m_l((Q_0)_{k,l} - (Q_0)_{j,l})^2$.
Then, there exist three positive constants c, c', c'', such that for any $1/m \leq \alpha_n L \leq 1/\log(n)$,

$$\frac{1}{m} \leq \beta \leq \beta(\alpha_n L) := \frac{K^3}{n} e^{4n\alpha_n L}$$

and

$$s^2 \geq c'' n/m,$$

with probability at least $1 - c/n^2$,

$$\mathrm{err}(\hat{G}, G) \leq e^{-c' s^2}.$$

In particular, since

$$s^2 = \frac{\alpha_n \min_{k\neq j} \sum_{l\in[K]} m_l((Q_0)_{k,l} - (Q_0)_{j,l})^2}{L} \geq \frac{\alpha_n m D^2}{L},$$

we get that with probability at least $1 - c/n^2$,

$$-\log\Big(\mathrm{err}(\hat{G}, G)\Big) = \Omega(m\alpha_n).$$

Presentation of the SDP-based clustering algorithm. In this Section, we present how we estimate the partition of the nodes $\hat{G}$ when communities are assigned using a Markovian dynamic. Our main result Theorem 2.2 shows that we are able to achieve

$$-\log\ \mathrm{err}(\hat{G}, G) = \Omega(n\alpha_n).$$

Stated otherwise, we get a misclassification error that decays exponentially fast with respect to $n\alpha_n$. We recover the convergence rate recently proved in Giraud and Verzelen [2019] in the standard SBM[4] when the size of the smallest cluster scales linearly with n like in our case. To reach this result, we use the SDP algorithm proposed by Giraud and Verzelen in Giraud and Verzelen [2019]. In the following, we expose how the method works.
Suppose the community of each node in the graph has been assigned. In all this subsection, all the communities are considered fixed. We denote X the adjacency matrix of the graph and we refer to Theorem 2.17 for the definition of $(m_k)_k$ and m. Giraud and Verzelen [2019] are interested in solving optimization problem similar to the following

$$\max_{B\in\mathcal{C}'} \langle X, B\rangle \ \text{ with} \tag{2.13}$$
$$\mathcal{C}' := \{B : \mathrm{PSD},\ B_{k,l} \geq 0,\ |B|_1 = \sum_k m_k^2\},$$

where PSD means that B is positive semidefinite and where $|\cdot|_1$ is the element-wise l_1 norm, namely the sum of the absolute values of all entries of a given matrix.

[4] See Theorem 2.17.

We remind that, dealing with two communities, when the values of the probability matrix Q are a constant p on the diagonal and another constant q off the diagonal with $p > q$, we are in the assortative case. In the assortative setting, optimization problems like (2.13) have been widely used to recover communities, see Chen and Xu [2016], Guédon and Vershynin [2014], Perry and Wein [2017], Hajek et al. [2016], Fei and Chen [2018]. Those SDP programs are trying to maximize the probability of connection between nodes belonging to the same community. Therefore, they cannot be used directly to solve community detection outside of the assortative framework.
Peng and Wei [2007] showed that any partition G of $[n]$ can be uniquely represented by a $n \times n$ matrix $B^* \in \mathbb{R}^{n\times n}$ defined by $\forall i, j \in [n]$,

$$B^*_{i,j} = \begin{cases} \dfrac{1}{m_k} & \text{if } i \text{ and } j \text{ belong to community } k \\ 0 & \text{otherwise.} \end{cases}$$

The set of such matrices B^* that can be built from a particular partition of $[n]$ in K groups is defined by

$$\mathcal{S} = \{B \in \mathbb{R}^{n\times n} \ : B^\top = B,\ B^2 = B,\ \mathrm{Tr}(B) = K, \\ B\mathbf{1} = \mathbf{1},\ B \geq 0\},$$

where $\mathbf{1} \in \mathbb{R}^n$ is the n-dimensional vector with all entries equal to one and where $B \geq 0$ means that all entries of B are nonnegative. Peng and Wei [2007] proved that solving the K-means problem

$$\mathrm{Crit}(G) = \sum_{k=1}^{K} \sum_{i\in G_k} \left\| X_{:,i} - \frac{1}{|G_k|} \sum_{j\in G_k} X_{:,j} \right\|^2,$$

is equivalent to

$$\max_{B\in\mathcal{S}} \langle XX^\top, B\rangle. \tag{2.14}$$

Writing B^* an optimal solution of (2.14), an optimal solution for the K-means problem is obtained by gathering indices $i, j \in [n]$ such that $B^*_{i,j} \neq 0$. The set $\mathcal{S}$ is not convex and the authors of Giraud and Verzelen [2019] propose the following relaxation of problem (2.14)

$$\hat{B} \in \arg\max_{B\in\mathcal{C}_\beta} \langle XX^\top, B\rangle \text{ with} \tag{2.15}$$
$$\mathcal{C}_\beta := \{B \in \mathbb{R}^{n\times n} \ : \text{ symmetric, } \mathrm{Tr}(B) = K, \\ B\mathbf{1} = \mathbf{1}, 0 \leq B \leq \beta\},$$

where $K/n \leq \beta \leq 1$. The constraint $B \leq \beta$ allows to deal with sparse graphs. Indeed, when $\alpha_n = o(\log(n)/n)$, solving (2.15) without this constraint will produce unbalanced partition.
At this step, we cannot ensure that $\hat{B}$ belongs to $\mathcal{S}$ and a final refinement is necessary to end up with a clustering of the nodes of the graph. This final rounding step is achieved by running a K-medoid algorithm on the rows of $\hat{B}$. Given a partition $\{G_1, \ldots, G_k\}$ of the n nodes of the graph into K communities, we define the related membership matrix $A \in \mathbb{R}^{n\times K}$ where $A_{i,k} = \mathbb{1}_{i\in G_k}$. Working on the rows of $\hat{B}$, a K-medoid algorithm tries to find efficiently a pair $(\hat{A}, \hat{M})$ with $\hat{A} \in \mathcal{A}_K$, $\hat{M} \in \mathbb{R}^{K\times n}$, $\mathrm{Rows}(\hat{M}) \subset \mathrm{Rows}(\hat{B})$ satisfying for some $\rho > 0$

$$|\hat{A}\hat{M} - \hat{B}|_1 \leq \rho \min_{A\in\mathcal{A}_K, \mathrm{Rows}(M)\subset\mathrm{Rows}(\hat{B})} |AM - \hat{B}|_1, \tag{2.16}$$

where A_K is the set of all possible membership matrices and $\mathrm{Rows}(\hat{B})$ the set of all rows of $\hat{B}$. The K-medoids algorithm proposed in Charikar et al. [2002] gives in polynomial time a pair $(\hat{A}, \hat{M})$ satisfying the inequality (2.16) with $\rho = 7$. From $\hat{A}$ we are able to define the final partition of the nodes of the graph by setting

$$\forall k \in [K], \quad \hat{G}_k = \{i \in [n] \ : \ \hat{A}_{i,k} = 1\}.$$

Remark.
As highlighted in Giraud and Verzelen [2019], the parameter β can not be computed since L is unknown. Verzelen and Giraud propose to set β to value $\hat{\beta} = \frac{K^3}{n} e^{2nd_X} \wedge 1$, where d_X denotes the density of the

graph. We end up with the Algorithm 1 to estimate the communities in the SBM.

Algorithm 1 Algorithm to estimate the partition of the nodes of the graph.

Data: Adjacency matrix X of a graph $G = (V, E)$, Number of communities K.

1: Compute the density of the graph $d_X = \frac{2|E|}{n(n-1)}$ and set $\hat{\beta} = \frac{K^3}{n} e^{2nd_X} \wedge 1$.
2: Find $\hat{B} \in \underset{B \in \mathcal{C}_{\hat{\beta}}}{\arg\max} \langle XX^\top, B \rangle$ (using for example the interior-point method).
3: Run the K-medoids algorithm from Charikar et al. [2002] on the rows of $\hat{B}$. Note $\hat{A} \in \{0,1\}^{n \times K}$ the membership matrix obtained.
4: Define $\forall k \in [K], \quad \hat{G}_k = \{i \in [n] \ : \ \hat{A}_{i,k} = 1\}$ and $\forall i \in [n], \quad \hat{C}_i = k$ where $k \in [K]$ is such that $\hat{A}_{i,k} = 1$.

2.9 Additional Experiments

2.9.1 Experiments with 2 communities

We test our algorithm on a toy example with $K = 2$ communities, $\alpha_n = 1$ and with the following matrices:

$$P = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix} \text{ and } Q_0 = \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.3 \end{bmatrix}. \tag{2.17}$$

The Figure 2.10 shows the evolution of the infinity norm of the difference between the true transition matrix P and our estimate $\hat{P}$ when the size of the graph is increasing. Those numerical results are consistent with Theorem 2.5: we recover the parametric convergence rate with our estimator of the transition matrix.

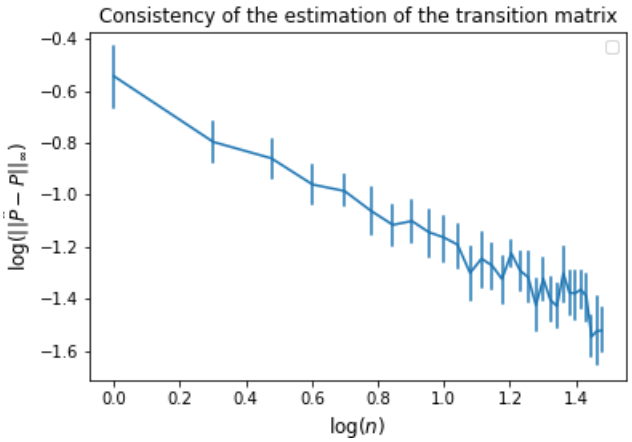


Figure 2.10: We plot the log of the infinity norm of the difference between the true transition matrix P and our estimate $\hat{P}$ according to the log of the number of nodes in the graph. For each point, the bar represents the standard deviation of the infinity norm error computed over thirty randomly generated graphs with the same number of nodes and using the matrices P and Q defined by (2.17).

With Figure 2.11, we shed light on the influence of the average degree of the nodes on the performance of our algorithm. We propose to compute the precision and the recall of the binary classification problem that we study when $K = 2$ defining

$$Q = \alpha \times Q_0,$$

where Q_0 is defined in (2.17) and α varies between 0.1 and 1 on a log scale. We remind that in a binary classification problem, the precision is the ratio between the number of examples labeled 1 that belong to class 1 and the number of examples labeled 1. The recall is the ratio between the number of examples labeled 1 that belong to class 1 and the number of examples that belong to class 1. In our context, those

definitions read as

$$\text{precision} = \frac{\sum_{i=1}^{n} \mathbb{1}\{\hat{C}_i = 1,\ C_i = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{\hat{C}_i = 1\}} \quad \text{and} \quad \text{recall} = \frac{\sum_{i=1}^{n} \mathbb{1}\{\hat{C}_i = 1,\ C_i = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{C_i = 1\}}.$$

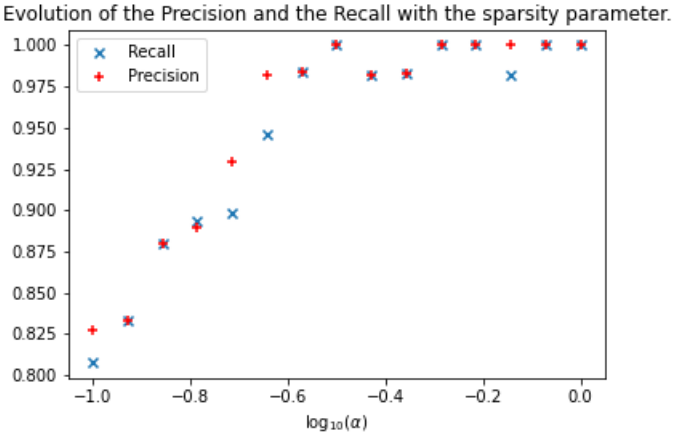


Figure 2.11: We plot the recall and the precision of the output of our algorithm with a graph sampled from SBM with a Markovian assignment of the communities using $n = 100$ nodes, a transition matrix P defined in (2.17) and a connectivity matrix $Q = \alpha Q_0$ where Q_0 is defined in (2.17) and α varies on a log scale between 0.1 and 1. We show the recall and the precision with respect to the $\log_{10}$ of the sparsity parameter α.

2.9.2 Experiments with 5 communities

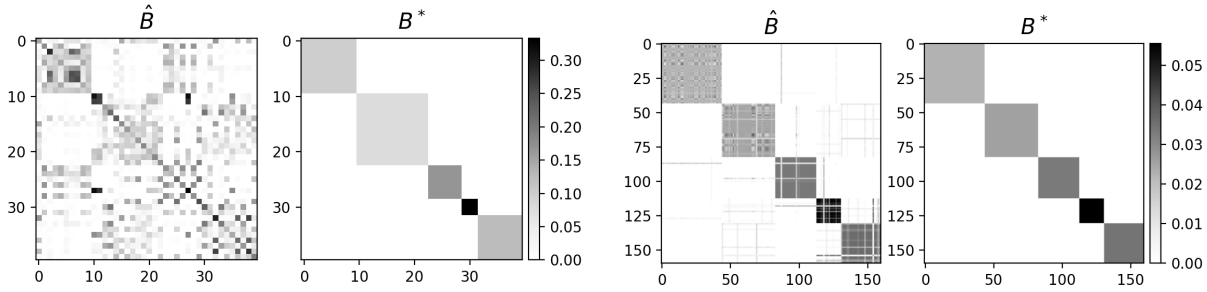
We test our algorithm on a toy example with $K = 5$ communities, with the transition matrix P and the connectivity matrix Q defined by

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.5 & 0.01 & 0.09 \\ 0.55 & 0.15 & 0.1 & 0.05 & 0.15 \\ 0.15 & 0.3 & 0.1 & 0.2 & 0.25 \\ 0.15 & 0.05 & 0.1 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.1 & 0.05 & 0.35 \end{bmatrix} \text{ and } Q = \begin{bmatrix} 0.6 & 0.1 & 0.15 & 0.1 & 0.2 \\ 0.2 & 0.5 & 0.35 & 0.1 & 0.4 \\ 0.4 & 0.15 & 0.6 & 0.25 & 0.05 \\ 0.4 & 0.1 & 0.1 & 0.2 & 0.55 \\ 0.3 & 0.35 & 0.2 & 0.1 & 0.7 \end{bmatrix}. \tag{2.18}$$

Sampling random graphs from SBM with Markovian assignment of the communities using the matrices (2.18), we see with Figure 2.12 that communities 3 and 4 have small sizes compared to the other clusters. For a graph sampled with a size equal to 40, Figure 2.12.a shows us that the SDP algorithm defined in Algorithm 1 is able to capture relevant information about the clustering of the nodes in communities $1,\ 2$ and 5. However, we see that using a number of nodes equal to 40 is not enough to distinguish nodes belonging to community 3 or 4. Figure 2.12.b proves that increasing the size of the graph (with $n = 160$) allows to solve this issue. One can easily guess that running a K-medoid algorithm on the rows of the matrix $\hat{B}$ plotted in Figure 2.12.b will lead to an accurate clustering of the nodes of the graph. Figure 2.13 shows that the $\log$ of the misclassification error decreases linearly with the size of the graph.

2.9.3 Other potential application on real data: The example of recommendation system

In this section, we give more details on another possible application of our model for recommendation system as mentioned in the introduction. Let us remind the framework of our example. We suppose



(a) $n = 40$. (b) $n = 160$.

Figure 2.12: We consider $K = 5$ communities and we order the nodes of the graph such that the true partition of the nodes is given by $G_1 = \{1, \dots, m_1\}, G_2 = \{m_1 + 1, \dots, m_1 + m_2\}, \dots, G_5 = \{\sum_{j=1}^{4} m_j + 1, \dots, n\}$. We generate random graphs from SBM with Markovian assignment of the communities using the transition matrix P and the connectivity matrix Q defined by 2.18. We plot the matrix B^* solution of Eq.(2.13) and its approximation $\hat{B}$ obtained by solving the SDP of Eq.(2.15). Thanks to the node ordering, the matrix B^* has a block diagonal structure where each entry of one block is equal to the inverse of the size of the associated cluster. Figure (a) allows us to compare the matrices B^* and $\hat{B}$ when the number of nodes in the graph is equal to 40 while Figure (b) deals with a graph of size 160.

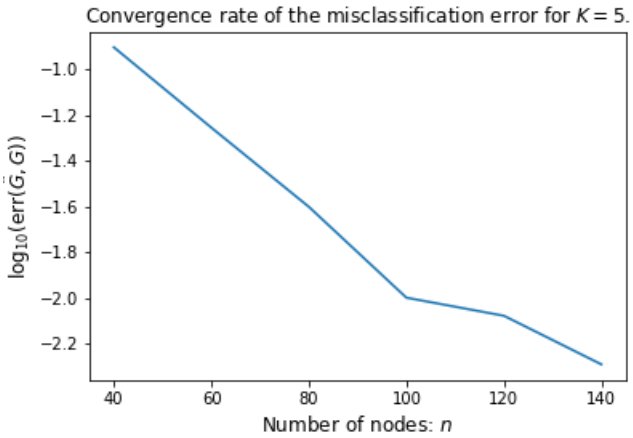


Figure 2.13: We consider $K = 5$ communities and we sample random graph from SBM with Markovian assignment of the communities using matrices defined in 2.18. We estimate the partition of the nodes of the graph using Algorithm 1. We plot the $\log$ of the misclassification error as a function of the size of the graphs sampled.

that we have access to the online purchases of different customers. For each of them, we know the dates and the product IDs of each of their purchases. Our goal is threefold: i) learn the category of product sold by each url ii) learn the purchasing behavior of each customer iii) use this information to suggest relevant new products to each customer. For each customer U, we have a network where nodes are product IDs (ordered by timestamp of purchase). We connect two products i and j if the ratio $W_{i,j}/\sqrt{w_i \times w_j}$ is larger than some threshold $\tau \in (0, 1)$, where $W_{i,j}$ is the number of clients in the dataset who bought both the products i and j, and w_i is the number of clients who bought the product i. We can proceed as follows.

1. Running our algorithm, we can infer the number of different categories of products bought by U using our heuristic from Section 2.6.2.

2. Then, we can learn both the category of each product and the transition matrix $\hat{P}$ which gives the purchasing behavior of U.

3. To recommend a new product to the client U, one can use the purchasing behaviour of the client V who shares the largest number of common purchased products with U.

Let us finally mention that the connection probabilities learned from client V for categories unseen so far by U could be used as an initialization of a stochastic bandit algorithm for recommendation of products for U.

Chapter 3

Markov Random Geometric Graphs: A growth model for temporal dynamic network

Chapter Abstract

In this chapter, we introduce the Markov Random Geometric Graphs: an extension of RGGs on the Euclidean Sphere where latent positions are sampled using an isotropic Markovian dynamic. We provide efficient algorithm to achieve non-parametric estimation of the connection function and of the Markov transition kernel in this model with theoretical guarantees. We stress the utility of this model by solving link prediction tasks. At the end of Section 3.8, we provide with Figure 3.15 a synthetic presentation of the estimation methods of this chapter.

Chapter Content

3.1 Introduction

In Random Geometric Graphs (RGG), nodes are sampled independently in latent space $\mathbb{R}^d$. Two nodes are connected if their distance is smaller than a threshold. A thorough probabilistic study of RGGs can be found in Penrose [2003]. RGGs have been widely studied recently due to their ability to provide a powerful modeling tool for networks with spatial structure. We can mention applications in bioinformatics Higham et al. [2008b] or analysis of social media Hoff et al. [2002]. One main feature is to uncover hidden representation of nodes using latent space and to model interactions by relative positions between latent points.
Furthermore, nodes interactions may evolve with time. In some applications, this evolution is given by the arrival of new nodes as in online collection growth Lo et al. [2017], online social network growth Backstrom et al. [2006], Jin et al. [2001], or outbreak modeling Ugander et al. [2012] for instance. The network is growing as more nodes are entering. Other time evolution modelings have been studied, we refer to Rossetti and Cazabet [2018] for a review.
A natural extension of RGG consists in accounting this time evolution. In Díaz et al. [2008], the expected length of connectivity and dis-connectivity periods of the Dynamic Random Geometric Graph is studied: each node choose at random an angle in $[0, 2\pi)$ and make a constant step size move in that direction. In Schott and Staples [2010], a random walk model for RGG on the hypercube is studied where at each time step a vertex is either appended or deleted from the graph. Their model falls into the class of Geometric Markovian Random Graphs that are generally defined in Clementi et al. [2009].
As far as we know, there is no extension of RGG to growth model for temporal dynamic networks. For the first time, we consider a Markovian dynamic on the latent space where the new latent point is drawn with respect to the latest latent point and some Markov kernel to be estimated.

Estimation of graphon in RGGs: the Euclidean sphere case Random graphs with latent space can be defined using a *graphon*, cf. Lovász [2012]. A graphon is a kernel function that defines edge distribution. In Tang et al. [2013], Tang and al. prove that spectral method can recover the matrix formed by graphon evaluated at latent points up to an orthogonal transformation, assuming that graphon is a positive definite kernel (PSD). Going further, algorithms have been designed to estimate graphons, as in Klopp et al. [2017] which provide sharp rates for the Stochastic Block Model (SBM). Recently, the paper De Castro et al. [2019] provides a non-parametric algorithm to estimate RGGs on Euclidean spheres, without PSD assumption.
We present here RGG on Euclidean sphere. Given n points $X_1, X_2, \dots, X_n$ on the Euclidean sphere $\mathbb{S}^{d-1}$, we set an edge between nodes i and j (where $i, j \in [n], i \neq j$) with independent probability $\mathbf{p}(\langle X_i, X_j \rangle)$. The unknown function $\mathbf{p} : [-1, 1] \to [0, 1]$ is called the *envelope function*. This RGG is a graphon model with a symmetric kernel W given by $W(x, y) = \mathbf{p}(\langle x, y \rangle)$. Once the latent points are given, independently draw the random undirected adjacency matrix A by

$$A_{i,j} \sim \mathrm{Ber}(\mathbf{p}(\langle X_i, X_j \rangle)), \quad i < j$$

with Bernoulli r.v. drawn independently (set zero on the diagonal and complete by symmetry), and set

$$T_n := \frac{1}{n} \left(\mathbf{p}(\langle X_i, X_j \rangle)\right)_{i,j \in [n]} \quad \text{and} \quad \widehat{T}_n := \frac{1}{n} A, \tag{3.1}$$

We do not observe the latent points and we have to estimate the envelope $\mathbf{p}$ from A only. A standard strategy is to remark that $\widehat{T}_n$ is a random perturbation of T_n and to dig into T_n to uncover $\mathbf{p}$.
One important feature of this model is that the interactions between nodes is depicted by a simple object: the envelope function $\mathbf{p}$. The envelope summarises how individuals connect each others given their latent positions. Standard examples Bubeck et al. [2016] are given by $\mathbf{p}_\tau(t) = \mathbb{1}_{\{t \geq \tau\}}$ where one connects two points as soon as their geodesic distance is below some threshold. The non-parametric estimation of $\mathbf{p}$ is given by De Castro et al. [2019] where the authors assume that latent points X_i are independently and uniformly distributed on the sphere, which will not be the case in this work.

A new growth model: the latent Markovian dynamic Consider RGGs where latent points are sampled with Markovian jumps, the Graphical Model under consideration can be found in Figure 3.1. Namely, we sample n points $X_1, X_2, \dots, X_n$ on the Euclidean sphere $\mathbb{S}^{d-1}$ using a Markovian dynamic. We start by sampling randomly X_1 on $\mathbb{S}^{d-1}$. Then, for any $i \in \{2, \dots, n\}$, we sample

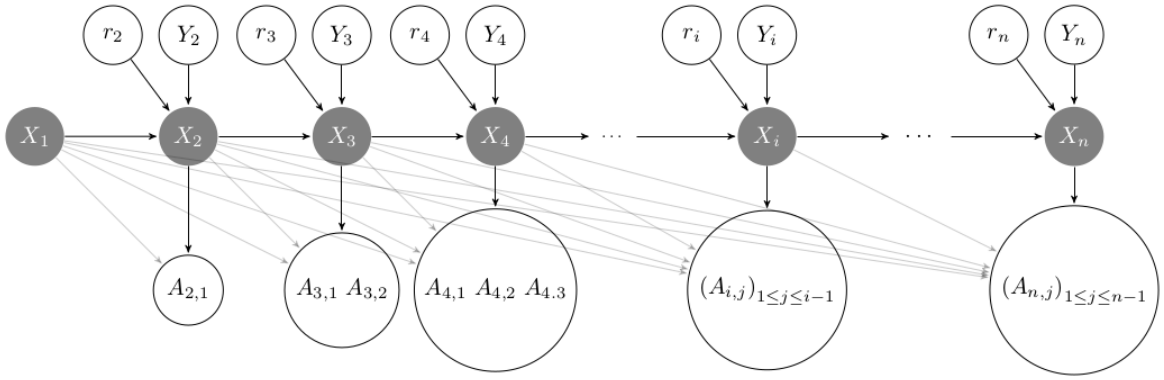
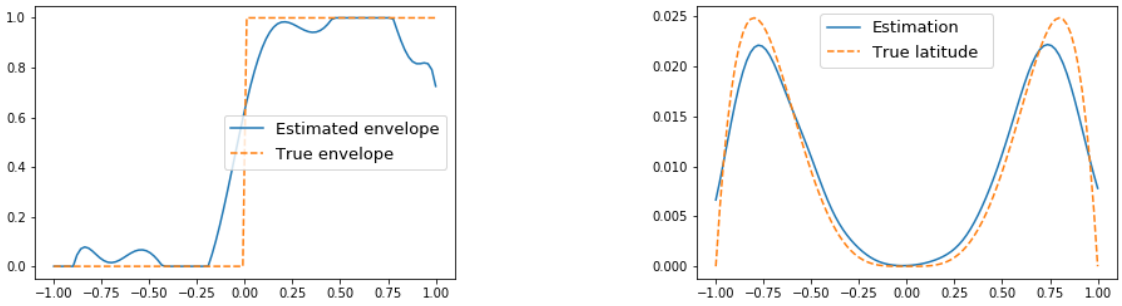


Figure 3.1: Graphical model of the MRGG model: Markovian dynamics on Euclidean sphere where we jump from X_k onto X_{k+1}. The Y_k encodes direction of jump while r_k encodes its distance, see (3.1).



(a) Envelope function

(b) Latitude function

Figure 3.2: Non-parametric estimation of envelope and latitude functions using algorithms of Sections 3.3 and 3.4. We built a graph of 1500 nodes sampled on the sphere $\mathbb{S}^2$ and using envelope $\mathbf{p}^{(1)}$ and latitude $f_{\mathcal{L}}^{(1)}$ (dot orange curves) defined in Section 3.6 by Eq.(3.12). The estimated envelope is thresholded to get a function in $[0,1]$ and the estimated latitude function is normalized with integral 1 (plain blue lines).

- a unit vector $Y_i \in \mathbb{S}^{d-1}$ uniformly, orthogonal to X_{i-1}.
- a real $r_i \in [-1,1]$ encoding the distance between X_{i-1} and X_i, see (3.2). r_i is sampled from a distribution $f_{\mathcal{L}} : [-1,1] \to [0,1]$, called the *latitude function*.

then X_i is defined by

$$X_i = r_i \times X_{i-1} + \sqrt{1-r_i^2} \times Y_i \,.$$

This dynamic can be pictured as follows. Consider that X_{i-1} is the north pole, then chose uniformly a direction (i.e., a longitude) and, in a independent manner, randomly move along the latitudes (the longitude being fixed by the previous step). The geodesic distance γ_i drawn on the latitudes satisfies

$$\gamma_i = \arccos(r_i)\,, \tag{3.2}$$

where random variable $r_i = \langle X_i, X_{i-1}\rangle$ has density $f_{\mathcal{L}}(r_i)$. The resulting model will be referred to as the Markov Random Geometric Graph (MRGG) and is described with Figure 3.1.

Temporal Dynamic Networks: MRGG estimation strategy. Seldom growth models exist for temporal dynamic network modeling, see Rossetti and Cazabet [2018] for a review. In our model, we add one

node at a time making a Markovian jump from the previous latent position. It results in

$$\text{the observation of } (A_{i,j})_{1\leq j\leq i-1} \text{ at time } T = i\,,$$

as pictured in Figure 3.1. Namely, we observe how a new node connects to the previous ones. For such dynamic, we aim at estimating the model, namely envelope $\mathbf{p}$ and respectively latitude $f_{\mathcal{L}}$. These functions capture in a simple function on $\Omega = [-1, 1]$ the range of interaction of nodes (represented by $\mathbf{p}$) and respectively the dynamic of the jumps in latent space (represented by $f_{\mathcal{L}}$), where, in abscissa Ω, values $r = \langle X_i, X_j\rangle$ near 1 corresponds to close point $X_i \simeq X_j$ while values close to -1 corresponds to antipodal points $X_i \simeq -X_j$. These functions may be non-parametric.
From snapshots of the graph at different time steps, can we recover envelope and latitude functions? We prove that it is possible under mild conditions on the Markovian dynamic of the latent points and our approach is summed up with Figure 3.3.

Fundamental result
Spectral convergence of $\widehat{T}_n$ under Markovian dynamic, see Section 3.3.1

$$\Downarrow$$

Guarantee for the recovery of:		Algorithm
(a) envelope $\mathbf{p}$, see (3.7)	$\leftrightarrow$	**SCCHEi**
(b) latent distances r_i, see (3.11)	$\leftrightarrow$	**HEiC** Araya and De Castro [2019]

Figure 3.3: Presentation of our method to recover the envelope and the latitude functions.

Define $\lambda(T_n) := (\lambda_1, \ldots, \lambda_n)$ and resp. $\lambda(\widehat{T}_n) := (\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$ the spectrum of T_n and resp. $\widehat{T}_n$, see (3.1). Building clusters from $\lambda(\widehat{T}_n)$, Algorithm 2 (SCCHEi) estimates the spectrum of envelope $\mathbf{p}$ while Algorithm 4 Araya and De Castro [2019] (HEiC, cf. Section 3.4.3) extracts d eigenvectors of $\widehat{T}_n$ to uncover the Gram matrix of the latent positions. Both can then be used to estimate the unknown functions of our model (cf. Figure 3.2).

Previous works. The latent space approach to model dynamics of network has already been studied in a large span of recent works. Most of them focus on block models with dynamic generalizations covering discrete dynamic evolution via hidden Markov models (cf. Matias and Miele [2017]) or continuous time analysis via extended Kalman filter (cf. Xu and Hero [2014]). Yang and Koeppl [2018] and Durante and Dunson [2014] use a Gamma Markov process allowing to model evolving mixed membership in graphs using respectively the Bernoulli Poisson link function and the logistic function to generate edges from the latent space representation. While the above mentioned papers consider community based random graphs with fixed size where edges and communities change through time, we focus on growing RGGs on Euclidean sphere where new nodes are added along time.
Non-parametric estimation of RGGs on $\mathbb{S}^{d-1}$ has been investigated in De Castro et al. [2019] with i.i.d. latent points. Estimation of latent point relative distances with HEiC Algorithm has been introduced in Araya and De Castro [2019] under i.i.d. latent points assumption. Phase transitions on the detection of geometry in RGGs (against Erdös Rényi alternatives) has been investigated in Bubeck et al. [2016].
For the first time, we introduce latitude function and non-parametric estimations of envelope and latitude using new results on kernel matrices concentration with dependent variables.

Outline. Section 3.2 presents important tools from Harmonic Analysis for this chapter. Sections 3.3 and 3.4 present the estimation method with new theoretical results under Markovian dynamic. These new results are random matrices operator norm control and resp. U-statistics control under Markovian dynamic, presented in Section 3.10.3 and resp. Section 3.10.2. The envelope *adaptive* estimate is built from a size constrained clustering (Algorithm 2) tuned by slope heuristic Eq.(3.8), and the latitude function estimate (cf. Section 3.4.1) is derived from estimates of latent distances r_i. Our method can handle random graphs with logarithmic growth node degree (i.e., new comer at time $T = n$ connects to $\mathcal{O}(\log n)$ previous nodes), referred to as *relatively sparse* models, see Section 3.5. Sections 3.6 and 3.7

investigate synthetic data experiments. We propose heuristics to solve link prediction problems and to test for a Markovian dynamic. In Section 3.8, we dig deeper into the analysis of our methods by studying their behaviour under model mispecification or under slow mixing conditions. We present final remarks and future research directions. At the end of Section 3.8, we provide with Figure 3.15 a synthetic presentation of the estimation methods of this chapter.
The remaining two sections of this chapter are dedicated to proofs of our main results.

Notations. Consider a dimension $d \geq 3$. Denote by $\|\cdot\|_2$ (resp. $\langle\cdot,\cdot\rangle$) the Euclidean norm (resp. inner product) on $\mathbb{R}^d$. Consider the d-dimensional sphere $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \ : \ \|x\|_2 = 1\}$ and denote by π the uniform probability measure on $\mathbb{S}^{d-1}$. For any matrix $M = (m_{i,j})_{i,j} \in \mathbb{R}^{D_1 \times D_2}$, we define $\|M\|_F^2 := \sum_{i=1}^{D_1}\sum_{j=1}^{D_2} |m_{i,j}|^2$ and the operator norm of M as $\|M\| := \sup_{x\in\mathbb{S}^{D_2-1}} \|Mx\|_2$. For two real valued sequences $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$, denote $u_n \underset{n\to\infty}{=} \mathcal{O}(v_n)$ if there exist $k_1 > 0$ and $n_0 \in \mathbb{N}$ such that $\forall n > n_0$, $|u_n| \leq k_1|v_n|$. For any $x, y \in \mathbb{R}$, $x \wedge y := \min(x, y)$ and $x \vee y := \max(x, y)$. Given two sequences x, y of reals–completing finite sequences by zeros–such that $\sum_i x_i^2 + y_i^2 < \infty$, we define the ℓ_2 rearrangement distance $\delta_2(x, y)$ as

$$\delta_2^2(x, y) := \inf_{\sigma\in\mathfrak{S}} \sum_i (x_i - y_{\sigma(i)})^2 ,$$

where $\mathfrak{S}$ is the set of permutations with finite support. This pseudo-distance is useful to compare two spectra.

3.2 Tools from Harmonic Analysis

3.2.1 Spectral decomposition of the kernel

One can associate with $W(x, y) = \mathbf{p}(\langle x, y\rangle)$ the integral operator

$$\mathbb{T}_W : L^2(\mathbb{S}^{d-1}) \to L^2(\mathbb{S}^{d-1}),$$

such that for any $g \in L^2(\mathbb{S}^{d-1})$,

$$\forall x \in \mathbb{S}^{d-1}, \quad (\mathbb{T}_W g)(x) = \int_{\mathbb{S}^{d-1}} g(y)\mathbf{p}(\langle x, y\rangle)\pi(dy),$$

where π is the uniform probability measure on $\mathbb{S}^{d-1}$. The operator $\mathbb{T}_W$ is Hilbert-Schmidt and it has a countable number of bounded eigenvalues λ_k^* with zero as only accumulation point. The eigenfunctions of $\mathbb{T}_W$ have the remarkable property that they do not depend on $\mathbf{p}$ (cf. Dai and Xu [2013] Lemma 1.2.3): they are given by the real Spherical Harmonics. We denote $\mathcal{H}_l$ the space of real Spherical Harmonics of degree l with dimension d_l and with orthonormal basis $(Y_{l,j})_{j\in[d_l]}$ where

$$d_l := \dim(\mathcal{H}_l) = \begin{cases} 1 & \text{if } l = 0 \\ d & \text{if } l = 1 \\ \binom{l+d-1}{l} - \binom{l+d-3}{l-2} & \text{otherwise.} \end{cases}$$

We end up with the following spectral decomposition

$$\mathbf{p}(\langle x, y\rangle) = \sum_{l\geq 0} p_l^* \sum_{1\leq j\leq d_l} Y_{l,j}(x)Y_{l,j}(y) = \sum_{k\geq 0} p_k^* c_k G_k^\beta(\langle x, y\rangle) , \tag{3.3}$$

where $\lambda(\mathbb{T}_W) = \{p_0^*, p_1^*, \ldots, p_1^*, \ldots, p_l^*, \ldots, p_l^*, \ldots\}$ meaning that each eigenvalue p_l^* has multiplicity d_l; and G_k^β is the Gegenbauer polynomial of degree k with parameter $\beta := \frac{d-2}{2}$ and $c_k := \frac{2k+d-2}{d-2}$ (cf. Section 3.2). Since $\mathbf{p}$ is bounded, one has $\mathbf{p} \in L^2((-1, 1), w_\beta)$ where the weight function w_β is defined by $w_\beta(t) := (1 - t^2)^{\beta-\frac{1}{2}}$ and

$$L^2((-1, 1), w_\beta) := \big\{g : [-1, 1] \to \mathbb{R} \ \big| \ \|g\|_2^2 := \int_{-1}^{1} |g(t)|^2 w_\beta(t)dt < +\infty\big\}.$$

$\mathbf{p}$ can be decomposed as $\mathbf{p} \equiv \sum_{k\geq 0} p_k^* c_k G_k^\beta$ and the Gegenbauer polynomials G_k^β are an orthogonal basis of $L^2((-1,1), w_\beta)$. The eigenvalues $(p_k^*)_{k\geq 0}$ of the envelope function can be computed numerically through the formula

$$\forall l \geq 0, \quad p_l^* = \left(\frac{c_l b_d}{d_l}\right) \int_{-1}^{1} p(t) G_l^\beta(t) w_\beta(t) dt,$$

where $b_d := \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{d}{2}-\frac{1}{2})}$ with Γ the Gamma function. Hence, it is possible to recover the envelope function $\mathbf{p}$ thanks to the identity

$$\mathbf{p} = \sum_{l\geq 0} \sqrt{d_l} p_l^* \frac{G_l^\beta}{\|G_l^\beta\|_{L^2([-1,1],w_\beta)}} = \sum_{l\geq 0} p_l^* c_l G_l^\beta. \tag{3.4}$$

3.2.2 Finite rank approximation of the kernel

We define for all $R \in \mathbb{N}$, $\widetilde{R} := \sum_{l=0}^{R} d_l$. We introduce for any resolution level $R \in \mathbb{N}$ the truncated graphon W_R which is obtained from W by keeping only the $\widetilde{R}$ first eigenvalues, that is

$$\forall x, y \in \mathbb{S}^{d-1}, \quad W_R(x,y) := \sum_{k=0}^{R} p_k^* \sum_{l=1}^{d_k} Y_{k,l}(x) Y_{k,l}(y).$$

Note that W_R is the best L^2-approximation of rank R of the kernel W. Similarly, we denote for all $t \in [0,1]$, $\mathbf{p}_R(t) = \sum_{k=0}^{R} p_k^* c_k G_k^\beta(t)$.
We will need in our proof the following result which states that fixing one variable and integrating with respect to the other one with the uniform probability measure on $\mathbb{S}^{d-1}$ gives $\|W - W_R\|_2^2$.

Lemma 3.1. *For any $x \in \mathbb{S}^{d-1}$,*

$$\mathbb{E}_{X\sim\pi}[(W - W_R)^2(x, X)] = \|W - W_R\|_2^2,$$

where π is the uniform measure on the $\mathbb{S}^{d-1}$.

Proof of Lemma 3.1.

$$\begin{aligned}
&\mathbb{E}_{X\sim\pi}[(W - W_R)^2(x, X)] \\
&= \int_y (W - W_R)^2(x,y)\pi(dy) \\
&= \int_y \left(\sum_{r>R} p_r^* \sum_{l=1}^{d_r} Y_{r,l}(x) Y_{r,l}(y)\right)^2 \pi(dy) \\
&= \int_y \sum_{r_1,r_2>R} p_{r_1}^* p_{r_2}^* \sum_{l_1=1}^{d_{r_1}} \sum_{l_2=1}^{d_{r_2}} Y_{r_1,l_1}(x) Y_{r_1,l_1}(y) Y_{r_2,l_2}(x) Y_{r_2,l_2}(y) \pi(dy) \\
&= \sum_{r_1,r_2>R} p_{r_1}^* p_{r_2}^* \sum_{l_1=1}^{d_{r_1}} \sum_{l_2=1}^{d_{r_2}} Y_{r_1,l_1}(x) Y_{r_2,l_2}(x) \int_y Y_{r_1,l_1}(y) Y_{r_2,l_2}(y) \pi(dy).
\end{aligned}$$

Since $\int_y Y_{r,l}(y) Y_{r',l'} \pi(dy)$ is 1 if $r = r'$ and $l = l'$ and 0 otherwise, we have that

$$\begin{aligned}
\mathbb{E}_{X\sim\pi}[(W - W_R)^2(x, X)] &= \sum_{r>R} (p_r^*)^2 \sum_{l=1}^{d_r} Y_{r,l}(x)^2 \\
&= \sum_{r>R} (p_r^*)^2 d_r \quad \text{(using [Dai and Xu, 2013, Eq.(1.2.9)])} \\
&= \|W - W_R\|_2^2.
\end{aligned}$$

□

3.2.3 Weighted Sobolev space

The space $Z^s_{w_\beta}((-1,1))$ with regularity $s > 0$ is defined as the set of functions $g = \sum_{k\geq 0} g^*_k c_k G^\beta_k \in L^2((-1,1), w_\beta)$ such that

$$\|g\|^*_{Z^s_{w_\beta}((-1,1))} := \left[\sum_{l=0}^{\infty} d_l |g^*_l|^2 \left(1 + (l(l+2\beta))^s\right)\right]^{1/2} < \infty.$$

3.3 Nonparametric estimation of the envelope function

3.3.1 Integral operator spectrum estimation with dependent variables

One key result is a new control of U-statistics with latent Markov variables that we prove with full details in Chapter 4. In this Chapter, we provide a short presentation is this result in Section 3.10.2. We consider a set of hypotheses on the Markov chain $(X_i)_{i\geq 1}$ ensuring that our concentration inequality for U-statistic from Chapter 4 holds. Namely, we work under the following assumption.

Assumption A The latitude function $f_{\mathcal{L}}$ is such that $\|f_{\mathcal{L}}\|_\infty < \infty$ and makes the chain $(X_i)_{i\geq 1}$ uniformly ergodic.

Under Assumption A, we prove in Section 3.9 that the unique stationary distribution of the Markov chain $(X_i)_{i\geq 1}$ is the uniform probability measure on $\mathbb{S}^{d-1}$ denoted π. Theorem 3.2 is a theoretical guarantee for a random matrix approximation of the spectrum of integral operator with **dependent** latent variables. Theorem 3.20 in Section 3.10.3 gives explicitly the constants hidden in the big O below which depend on the absolute spectral gap of the Markov chain $(X_i)_{i\geq 1}$ (cf. Definition A.10).

Theorem 3.2. *We consider that Assumption A holds and we assume the envelope* $\mathbf{p}$ *has regularity* $s > 0$*. Then, it holds*

$$\mathbb{E}\left[\delta_2^2(\lambda(\mathbb{T}_W), \lambda(T_n))\right] = \quad \mathcal{O}\left(\left[\frac{n}{\log^2(n)}\right]^{-\frac{2s}{2s+d-1}}\right).$$

Using this preliminary result and the near optimal error bound for the operator norm of random matrices from Bandeira and van Handel [2016] we obtain

$$\mathbb{E}\left[\delta_2^2(\lambda(\mathbb{T}_W), \lambda^{R_{opt}}(\widehat{T}_n))\right] = \quad \mathcal{O}\left(\left[\frac{n}{\log^2(n)}\right]^{-\frac{2s}{2s+d-1}}\right),$$

with $\lambda^{R_{opt}}(\widehat{T}_n) = (\hat{\lambda}_1, \dots, \hat{\lambda}_{\widetilde{R}_{opt}}, 0, 0, \dots)$ *and* $R_{opt} = \lfloor (n/\log^2(n))^{\frac{1}{2s+d-1}} \rfloor$*.* $\hat{\lambda}_1, \dots, \hat{\lambda}_n$ *are the eigenvalues of* $\widehat{T}_n$ *sorted in decreasing order of magnitude.*

Remark. In Theorem 3.2 and Theorem 3.8, note that we recover, up to a log factor, the *minimax rate of non-parametric estimation* of s-regular functions on a space of (Riemannian) dimension $d-1$. Even with i.i.d. latent variables, it is still an open question to know if this rate is the minimax rate of non-parametric estimation of RGGs.
Eq.(3.3) shows that one could use an approximation of $(p^*_k)_{k\geq 1}$ to estimate the envelope $\mathbf{p}$ and Theorem 3.2 states we can recover $(p^*_k)_{k\geq 1}$ up to a permutation. In most cases, the problem of finding such a permutation is NP-hard and we introduce in the next section an efficient algorithm to fix this issue.

3.3.2 Size Constrained Clustering Algorithm

Note the spectrum of $\mathbb{T}_W$ is given by $(p^*_l)_{l\geq 0}$ where p^*_l has multiplicity d_l. In order to recover envelope $\mathbf{p}$, we build clusters from eigenvalues of $\widehat{T}_n$ while respecting the dimension d_l of each eigen-space of $\mathbb{T}_W$. In De Castro et al. [2019], an algorithm is proposed testing all permutations of $\{0, \dots, R\}$ for a given maximal resolution R. To bypass the high computational cost of such approach, we propose an efficient method based on the tree built from *Hierarchical Agglomerative Clustering* (HAC). In the following, for any $\nu_1, \dots, \nu_n \in \mathbb{R}$, we denote by $\mathrm{HAC}(\{\nu_1, , \dots, \nu_n\}, d_c)$ the tree built by a HAC on the real values $\nu_1, \dots, \nu_n$ using the complete linkage function d_c defined by $\forall A, B \subset \mathbb{R}$, $d_c(A,B) = \max_{a\in A} \max_{b\in B} \|a-b\|_2$. Algorithm 2 describes our approach.

Algorithm 2 Size Constrained Clustering for Harmonic Eigenvalues (SCCHEi).

Data: Resolution R, matrix $\widehat{T}_n = \frac{1}{n}A$, dimensions $(d_k)_{k=0}^R$.

1: Let $\hat{\lambda}_1, \dots, \hat{\lambda}_n$ be the eigenvalues of $\widehat{T}_n$ sorted in decreasing order of magnitude.
2: Set $\mathcal{P} := \{\hat{\lambda}_1, \dots, \hat{\lambda}_{\widetilde{R}}\}$ and $dims = [d_0, d_1, \dots, d_R]$.
3: **while** All eigenvalues in $\mathcal{P}$ are not clustered **do**
4: $tree \leftarrow$ HAC(nonclustered eigenvalues in $\mathcal{P}$, d_c)
5: **for** $d \in dims$ **do**
6: Search for a cluster of size d in $tree$ as close as possible to the root.
7: **if** such a cluster $\mathcal{C}_d$ exists **then** Update$(dims, tree, \mathcal{C}_d, d)$.
8: **end for**
9: **for** $d \in dims$ **do**
10: Search for the group $\mathcal{C}$ in $tree$ with a size larger than d and as close as possible to d.
11: **if** such a group exists **then** Update$(dims, tree, \mathcal{C}, d)$ **else** Go to line 3.
12: **end for**
13: **end while**

Return: $\mathcal{C}_{d_0}, \dots, \mathcal{C}_{d_R}, \{\hat{\lambda}_{\widetilde{R}+1}, \dots, \hat{\lambda}_n\}$

Algorithm 3 Update$(dims, tree, \mathcal{C}, d)$.

1: Save the subset $\mathcal{C}_d$ consisting of the d eigenvalues in $\mathcal{C}$ with the largest absolute values.
2: Delete from $tree$ all occurrences to eigenvalues in $\mathcal{C}_d$ and delete d from $dims$.

Given some resolution level $R \in \mathbb{N}$, our estimator $\widehat{\mathbf{p}}_R$ of the envelope function $\mathbf{p}$ is obtained from the clustering of the eigenvalues obtained by the SCCHEi algorithm as follows

$$\widehat{\mathbf{p}}_R : t \mapsto \sum_{k=0}^{R} \widehat{p}_k c_k G_k^{\beta}(t) \quad \text{where} \quad \forall k \in \{0, \dots, R\}, \quad \widehat{p}_k := \frac{1}{d_k} \sum_{\lambda \in \mathcal{C}_{d_k}} \lambda. \tag{3.5}$$

3.3.3 Theoretical guarantees

Let us recall that for any resolution level $R \geq 0$,

$$\lambda(\mathbb{T}_{W_R}) = (\lambda_1^*, \dots, \lambda_{\widetilde{R}}^*, 0, 0, \dots) \text{ and } \lambda^R(\widehat{T}_n) = (\hat{\lambda}_1, \dots, \hat{\lambda}_{\widetilde{R}}, 0, 0, \dots)$$

where $\hat{\lambda}_1, \dots, \hat{\lambda}_n$ are the eigenvalues of $\widehat{T}_n$ sorted in decreasing order of magnitude. We order the eigenvalues $\hat{\lambda}_1, \dots, \hat{\lambda}_{\widetilde{R}}$ and in the following we consider that $\lambda^R(\widehat{T}_n)_1 \geq \dots \geq \lambda^R(\widehat{T}_n)_{\widetilde{R}}$.

Theorem 3.3. *Let us consider some resolution level $R \in \mathbb{N}$. We keep the assumptions of Theorem 1. We recall that we consider $\lambda^R(\widehat{T}_n)_1 \geq \dots \geq \lambda^R(\widehat{T}_n)_{\widetilde{R}}$.*
Then for n large enough, the clusters $\mathcal{C}_{d_0}, \dots, \mathcal{C}_{d_R}$ obtained from the SCCHEi algorithm satisfy

$$\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) = \sum_{k=0}^{R} \sum_{\hat{\lambda} \in \mathcal{C}_{d_k}} (\hat{\lambda} - p_k^*)^2.$$

Proof of Theorem 3.3. Let us denote

$$\Delta^G = \min_{0 \leq k \neq l \leq R,\ p_k^* \neq p_l^*} |p_k^* - p_l^*| \wedge \min_{0 \leq k \leq R,\ p_k^* \neq 0} |p_k^*| > 0.$$

For any $g \in (0, \frac{\Delta^G}{4})$, the proof of Theorem 3.2 (cf. Section 3.10.3) ensures that for n large enough it holds

$$\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) \leq g^2. \tag{3.6}$$

Let us recall that

$$\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) = \inf_{\sigma \in \mathfrak{S}} \sum_{i \geq 1} \left(\lambda(\mathbb{T}_{W_R})_{\sigma(i)} - \lambda^R(\widehat{T}_n)_i\right)^2.$$

The proof of Theorem 3.3 relies on the following two Lemmas. The proofs of these Lemmas are postponed to Section 3.10.1.

Lemma 3.4. *We keep the assumptions of Theorem 3.3. Then, for n large enough for Eq.(3.6) to hold, one can choose a permutation σ^* such that*

- $\sigma^*(\{1,\dots,\widetilde{R}\}) = \{1,\dots,\widetilde{R}\}$.
- $\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) = \sum_{i=1}^{\widetilde{R}} (\lambda(\mathbb{T}_{W_R})_{\sigma^*(i)} - \lambda^R(\widehat{T}_n)_i)^2$.

Moreover, the function f^* given by

$$\begin{aligned} f^* : \{1,\dots,\widetilde{R}\} &\to \{p_k^*,\ 0 \le k \le R\} \\ i &\mapsto \lambda(\mathbb{T}_{W_R})_{\sigma^*(i)}, \end{aligned}$$

is non-increasing.

Lemma 3.5. *We keep the assumptions and notations of Lemma 3.4. A clustering $\left(\widehat{\mathcal{C}}_{d_k}\right)_{0\le k\le R}$ at depth R in the tree of the HAC algorithm applied to $\mathcal{P} := \{\lambda^R(\widehat{T}_n)_1,\dots,\lambda^R(\widehat{T}_n)_{\widetilde{R}}\}$ is said to be of type $(\mathcal{S})$ if it satisfies:*

$$\begin{aligned} \widehat{\mathcal{C}}_{d_0} &\subset \{\lambda^R(\widehat{T}_n)_i \mid 1 \le i \le \widetilde{R},\ f^*(i) = p_0^*\}, \quad |\widehat{\mathcal{C}}_{d_0}| = d_0, \\ \widehat{\mathcal{C}}_{d_1} &\subset \{\lambda^R(\widehat{T}_n)_i \mid 1 \le i \le \widetilde{R},\ f^*(i) = p_1^*\}, \quad |\widehat{\mathcal{C}}_{d_1}| = d_1, \\ &\dots \\ \widehat{\mathcal{C}}_{d_R} &\subset \{\lambda^R(\widehat{T}_n)_i \mid 1 \le i \le \widetilde{R},\ f^*(i) = p_R^*\}, \quad |\widehat{\mathcal{C}}_{d_R}| = d_R. \end{aligned}$$

Then the HAC algorithm with complete linkage applied to $\mathcal{P}$ reaches (after $\widetilde{R}-R-1$ iterations) a state $\left(\widehat{\mathcal{C}}_{d_k}\right)_{0\le k\le R}$ of type $(\mathcal{S})$. As a consequence, the SCCHEi algorithm returns the clusters $\mathcal{C}_{d_0} = \widehat{\mathcal{C}}_{d_0},\dots,\mathcal{C}_{d_R} = \widehat{\mathcal{C}}_{d_R}$.

Theorem 3.3 directly follows from the conclusion of Lemma 3.5 since we get that

$$\begin{aligned} \sum_{k=0}^{R} \sum_{\hat{\lambda}\in\mathcal{C}_{d_k}} (\hat{\lambda} - p_k^*)^2 &= \sum_{i=1}^{\widetilde{R}} (\lambda^R(\widehat{T}_n)_i - f^*(i))^2 = \sum_{i=1}^{\widetilde{R}} (\lambda^R(\widehat{T}_n)_i - \lambda(\mathbb{T}_{W_R})_{\sigma^*(i)})^2 \\ &= \delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)), \end{aligned}$$

where the first equality comes from the conclusion of Lemma 3.5, the second one comes from the definition of f^* from Lemma 3.4 and the last one comes from the choice of σ^* from Lemma 3.4. □

Theorem 3.3 ensures that under appropriate conditions, the SCCHEi leads to a clustering of the eigenvalues of the adjacency matrix that achieves the δ_2 distance between $\lambda(\mathbb{T}_{W_R})$ and $\lambda^R(\widehat{T}_n)$. Nevertheless, this is not a sufficient condition to ensure that the L^2 error between the true envelope function and our plug-in estimator (cf. Eq.(3.5)) goes to 0 has $n \to +\infty$. This is due to identifiability issues coming from the δ_2 metric. This was already mentioned in [De Castro et al., 2019, Section 3.6], where the authors present the following example. Consider the case $d = 3$, which implies $\beta = 1/2$, $d_k = 2k+1$, $c_k = 2k+1$. For $\mu > 0$, let

$$\begin{aligned} \mathbf{p}_a &= \frac{1}{2}c_0 G_0^\beta + \mu c_1 G_1^\beta + 0 \times c_2 G_2^\beta + 0 \times c_3 G_3^\beta + \mu c_4 G_4^\beta \\ \mathbf{p}_b &= \frac{1}{2}c_0 G_0^\beta + 0 \times c_1 G_1^\beta + \mu c_2 G_2^\beta + \mu c_3 G_3^\beta + 0 \times c_4 G_4^\beta \end{aligned}$$

Then the associated spectrum are

$$\begin{aligned} \lambda_a^* &= (1/2, \underbrace{\mu,\mu,\mu}_{3}, \underbrace{0,0,0,0,0}_{5}, \underbrace{0,0,0,0,0,0,0}_{7}, \underbrace{\mu,\mu,\mu,\mu,\mu,\mu,\mu,\mu,\mu}_{9}) \\ \lambda_b^* &= (1/2, \underbrace{0,0,0}_{3}, \underbrace{\mu,\mu,\mu,\mu,\mu}_{5}, \underbrace{\mu,\mu,\mu,\mu,\mu,\mu,\mu}_{7}, \underbrace{0,0,0,0,0,0,0,0,0}_{9}) \end{aligned}$$

which are indistinguishable in δ_2 metric, although $\|\mathbf{p}_a - \mathbf{p}_b\|_2 = \mu\sqrt{24}$.
Nevertheless, we can obtain a theoretical guarantee on the L^2 error between the true envelope function and our plug-in estimate using Theorem 3.3 if we consider additional conditions on the eigenvalues $(p_k^*)_{k\geq 0}$.

Theorem 3.6. *Assume that the envelope function* $\mathbf{p}$ *is polynomial of degree* $D \in \mathbb{N}$, *i.e.,* $p_k^* = 0$ *for any* $k > D$ *and* $p_D^* \neq 0$. *Assume also that all nonzeros* p_k^* *for* $k \in \{0, \dots, D\}$ *are distinct and that* $R \geq D$. *Then for* n *large enough it holds with probability at least* $1 - n^{-8}$,

$$\|\widehat{\mathbf{p}}_R - \mathbf{p}\|_2^2 \leq c\frac{\widetilde{R}}{n}\ln(n),$$

where $c > 0$ *is a universal numerical constant.*

Remarks.

- The question of whether the problem of estimating $\mathbf{p}$ is NP-hard was still completely open. Theorem 3.6 brings a first partial answer to this question by showing that $\mathbf{p}$ can be estimated in polynomial time in the case where $\mathbf{p}$ is a polynomial with all non-zero eigenvalues distinct.
- The proof of Theorem 3.6 is strictly analogous to the one of [De Castro et al., 2019, Proposition 9]. In a nutshell, considering that the envelope function $\mathbf{p}$ is a polynomial with all non-zeros eigenvalues p_k^* distinct ensures that (since $R \geq D$)

  $$\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) = \delta_2^2(\lambda(\mathbb{T}_W), \lambda^R(\widehat{T}_n)),$$

  which coincides with the L^2 norm of the difference between $\mathbf{p}$ and its estimate

  $$\widehat{\mathbf{p}}_{opt,R} := \sum_{k=0}^{R} \widehat{p}_{opt,k} c_k G_k^\beta \quad \text{with} \quad \widehat{p}_{opt,k} := \frac{1}{d_k} \sum_{i \in (\sigma^*)^{-1}([\widetilde{k}+1, \widetilde{k+1}])} \lambda^R(\widehat{T}_n)_i,$$

  where σ^* is a permutation as defined in Lemma 3.4. Since we proved that for n large enough, the clusters returned by the SCCHEi algorithm correspond to an allocation given by f^*, we deduce that the L^2 norm between $\mathbf{p}$ and our plug-in estimate $\widehat{\mathbf{p}}_R$ is equal to the δ_2 distance between spectra. The result then comes directly using Theorem 3.2.

3.3.4 Adaptation: Slope heuristic as model selection of Resolution

A data-driven choice of model size R can be done by *slope heuristic*, see Arlot [2019] for a nice review. One main idea of slope heuristic is to penalize the empirical risk by $\kappa\,\mathrm{pen}(\widetilde{R})$ and to calibrate $\kappa > 0$. If the sequence $(\mathrm{pen}(\widetilde{R}))_{\widetilde{R}}$ is equivalent to the sequence of variances of the population risk of empirical risk minimizer (ERM) as model size $\widetilde{R}$ grows, then, penalizing the empirical risk (as done in Eq.(3.8)), one may ultimately uncover an empirical version of the U-shaped curve of the population risk. Hence, minimizing it, one builds a model size $\hat{R}$ that balances between bias (*under-fitting* regime) and variance (*over-fitting* regime). First, note that empirical risk is given by the intra-class variance below.

Definition 3.7. For any output $(\mathcal{C}_{d_0}, \dots, \mathcal{C}_{d_R}, \Lambda)$ of the Algorithm SCCHEi, the thresholded intra-class variance is defined by

$$\mathcal{I}_R := \frac{1}{n}\left[\sum_{k=0}^{R}\sum_{\lambda\in\mathcal{C}_{d_k}}\left(\lambda - \frac{1}{d_k}\sum_{\lambda'\in\mathcal{C}_{d_k}}\lambda'\right)^2 + \sum_{\lambda\in\Lambda}\lambda^2\right],$$

and the estimations $(\hat{p}_k)_{k\geq 0}$ of the eigenvalues $(p_k^*)_{k\geq 0}$ is given by

$$\forall k \in \mathbb{N}, \quad \hat{p}_k = \begin{cases} \frac{1}{d_k}\sum_{\lambda\in\mathcal{C}_{d_k}}\lambda & \text{if } k \in \{0, \dots, \hat{R}\} \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

Second, as underlined in the proof of Theorem 3.2 (see Theorem 3.20 in Section 3.10.3), the estimator's variance of our estimator scales linearly in $\widetilde{R}$.

Hence, we apply Algorithm SCCHEi for R varying from 0 to $R_{\max}$ (with $R_{\max} := \max\{R \geq 0 \ : \ \widetilde{R} \leq n\}$) to compute the *thresholded intra-class variance* $\mathcal{I}_R$ (see Definition 3.7) and given some $\kappa > 0$, we select

$$R(\kappa) \in \underset{R\in\{0,\dots,R_{max}\}}{\arg\min} \left\{ \mathcal{I}_R + \kappa \frac{\widetilde{R}}{n} \right\}. \tag{3.8}$$

The hyper-parameter κ controlling the bias-variance trade-off is set to $2\kappa_0$ where κ_0 is the value of $\kappa > 0$ leading to the "*largest jump*" of the function $\kappa \mapsto R(\kappa)$. Once $\hat{R} := R(2\kappa_0)$ has been computed, we approximate the envelope function $\mathbf{p}$ using Eq.(3.7) (see Eq.(3.4) for the closed form). We denote this estimator $\widehat{\mathbf{p}}$ and with the notations of Eq.(3.5) it holds $\widehat{\mathbf{p}} = \widehat{\mathbf{p}}_{\widehat{R}}$.
We propose a detailed analysis of the slope heuristic on simulated data using $d = 3$, the envelope function $\mathbf{p}^{(1)}$ and the latitude function $f_{\mathcal{L}}^{(1)}$ presented in Eq.(3.12). We recall that $R(\kappa)$ represents the optimal value of R to minimize the bias-variance decomposition defined by Eq.(3.8) for a given hyperparameter κ. Figure 3.4 shows the evolution of $\widetilde{R}(\kappa)$ with respect to κ which is sampled on a logscale. $\widetilde{R}(\kappa)$ is the dimension of the space of Spherical Harmonics with degree at most $R(\kappa)$. Our slope heuristic consists in choosing the value κ_0 leading to the larger jump of the function $\kappa \mapsto \widetilde{R}(\kappa)$. In our case, Figure 3.4 shows that $\kappa_0 = 10^{-3.9}$. As described in Section 3.3.2, the resolution level $\hat{R}$ selected to cluster the eigenvalues of the matrix $\widehat{T}_n$ is given by $R(2\kappa_0)$.

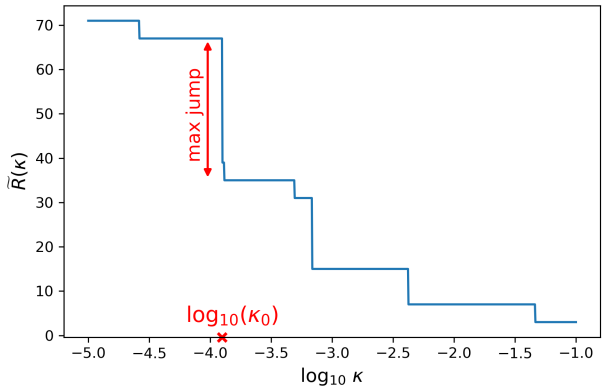


Figure 3.4: We sample the parameter κ on a logscale between 10^{-5} and 10^{-1} and we compute the corresponding $R(\kappa)$ defined in Eq.(3.8). We plot the values of $\widetilde{R}(\kappa)$ with respect to κ. The larger jump allows us to define κ_0.

These results can be reproduced using the notebook *Experiments*[1].

3.4 Nonparametric estimation of the latitude function

3.4.1 Our approach to estimate the latitude function in a nutshell

In Theorem 3.8 (see below), we show that we are able to estimate consistently the pairwise distances encoded by the Gram matrix G^* where

$$G^* := \frac{1}{n} \left(\langle X_i, X_j \rangle \right)_{i,j\in[n]}.$$

Taking the diagonal just above the main diagonal (referred to as *superdiagonal*) of $\widehat{G}$ - an estimate of the matrix G to be specified - we get estimates of the i.i.d. random variables $(\langle X_i, X_{i-1} \rangle)_{2\leq i\leq n} = (r_i)_{2\leq i\leq n}$ sampled from $f_{\mathcal{L}}$. Using $(\hat{r}_i)_{2\leq i\leq n}$ the superdiagonal of $n\widehat{G}$, we can build a kernel density estimator of the latitude function $f_{\mathcal{L}}$. In the following, we describe the algorithm used to build our estimator $\hat{G}$ with theoretical guarantees.

3.4.2 Spectral gap condition and Gram matrix estimation

The Gegenbauer polynomial of degree one is defined by $G_1^\beta(t) = 2\beta t, \ \forall t \in [-1, 1]$. As a consequence, using the *addition theorem* (cf. [Dai and Xu, 2013, Lem.1.2.3 and Thm.1.2.6]), the Gram matrix G^* is

[1] https://github.com/quentin-duchemin/Markovian-random-geometric-graph

related to the Gegenbauer polynomial of degree one. More precisely, for any $i, j \in [n]$ it holds

$$G^*_{i,j} = \frac{1}{2\beta n} G_1^{\beta}(\langle X_i, X_j \rangle) = \frac{1}{nd} \sum_{k=1}^{d} Y_{1,k}(X_i) Y_{1,k}(X_j). \tag{3.9}$$

Denoting for all $k \in [d]$ $v_k^* := \frac{1}{\sqrt{n}} \left(Y_{1,k}(X_1), \ldots, Y_{1,k}(X_n)\right) \in \mathbb{R}^n$, and $V^* = (v_1^*, \ldots, v_d^*) \in \mathbb{R}^{n \times d}$, Eq.(3.9) becomes

$$G^* := \frac{1}{d} V^* (V^*)^\top.$$

We will prove that for n large enough there exists a matrix $\widehat{V} \in \mathbb{R}^{n \times d}$ where each column is an eigenvector of $\widehat{T}_n$, such that $\widehat{G} := \frac{1}{d} \widehat{V} \widehat{V}^\top$ approximates G^* well, in the sense that the Frobenius norm $\|G^* - \widehat{G}\|_F$ converges to 0. To choose the d eigenvectors of the matrix $\widehat{T}_n$ that we will use to build the matrix $\widehat{V}$, we need the following spectral gap condition

$$\Delta^* := \min_{k \in \mathbb{N},\, k \neq 1} |p_1^* - p_k^*| > 0. \tag{3.10}$$

This condition will allow us to apply Davis-Kahan type inequalities.
Now, thanks to Theorem 3.2, we know that the spectrum of the matrix $\widehat{T}_n$ converges towards the spectrum of the integral operator $\mathbb{T}_W$. Then, based on Eq.(3.9), one can naturally think that extracting the d eigenvectors of the matrix $\widehat{T}_n$ related with the eigenvalues that converge towards p_1^*, we can approximate the Gram matrix G^* of the latent positions. Theorem 3.8 proves that the latter intuition is true with high probability under the spectral gap condition (3.10). The algorithm HEiC Araya and De Castro [2019] (presented in Section 3.4.3) aims at identifying the above mentioned d eigenvectors of the matrix $\widehat{T}_n$ to build our estimate of the Gram matrix G^*.

Theorem 3.8. *We consider that Assumption A holds, we assume $\Delta^* > 0$, and we assume that graphon W has regularity $s > 0$. We denote $\widehat{V} \in \mathbb{R}^{n \times d}$ the d eigenvectors of the matrix $\widehat{T}_n$ associated with the eigenvalues returned by the algorithm HEiC and we define $\widehat{G} := \frac{1}{d} \widehat{V} \widehat{V}^\top$. Then for n large enough and for some constant $D > 0$, it holds with probability at least $1 - 5/n^2$,*

$$\|G^* - \widehat{G}\|_F \leq D \left(\frac{n}{\log^2(n)} \right)^{\frac{-s}{2s+d-1}}. \tag{3.11}$$

Based on Theorem 3.8, we propose a kernel density approach to estimate the latitude function $f_{\mathcal{L}}$ based on the super-diagonal of the matrix $\widehat{G}$, namely $\left(\hat{r}_i := n \widehat{G}_{i-1,i}\right)_{i \in \{2, \ldots, n\}}$. In the following, we denote $\hat{f}_{\mathcal{L}}$ this estimator.

3.4.3 Harmonic EigenCluster(HEiC)

The HEiC algorithm was first introduced by Araya and De Castro [2019] and allows us to extract d eigenvectors from the matrix $\widehat{T}_n$ to compute our estimate $\widehat{G}$ of the Gram matrix G^*. Let us first define for a given set of indices $i_1, \ldots, i_d \in [n]$

$$\mathrm{Gap}_1(\widehat{T}_n; i_1, \ldots, i_d) := \min_{i \notin \{i_1, \ldots, i_d\}} \max_{j \in \{i_1, \ldots, i_d\}} |\hat{\lambda}_i - \hat{\lambda}_j|.$$

The HEiC procedure is presented in the following algorithm.

3.5 Relatively Sparse Regime

Although we deal so far with the so-called *dense* regime (i.e. when the expected number of neighbors of each node scales linearly with n), our results may be generalized to the *relatively sparse* model connecting nodes i and j with probability $W(X_i, X_j) = \zeta_n \mathbf{p}(\langle X_i, X_j \rangle)$ where $\zeta_n \in (0, 1]$ satisfies $\liminf_n \zeta_n n / \log n \geq Z$ for some universal constant $Z > 0$.
In the relatively sparse model, one can show following the proof of Theorem 3.2 that the resolution

Algorithm 4 Harmonic EigenCluster(HEiC) algorithm.

Data: Adjacency matrix A. Dimension d.

1: $(\hat{\lambda}_1^{sort}, \dots, \hat{\lambda}_n^{sort}) \leftarrow$ eigenvalues of $\widehat{T}_n$ sorted in decreasing order.
2: $\Lambda_1 \leftarrow \{\hat{\lambda}_1^{sort}, \dots, \hat{\lambda}_d^{sort}\}$.
3: Initialize $i = 2$ and gap$= \mathrm{Gap}_1(\widehat{T}_n; 1, 2, \dots, d)$.
4: **while** $i \leq n - d + 1$ **do**
5: **if** $\mathrm{Gap}_1(\widehat{T}_n; i, i+1, \dots, i+d-1) >$ gap **then**
6: $\Lambda_1 \leftarrow \{\hat{\lambda}_i^{sort}, \dots, \hat{\lambda}_{i+d-1}^{sort}\}$
7: **end if**
8: $i = i + 1$
9: **end while**

Return: Λ_1, gap.

should be chosen as $\widehat{R} = \left(\frac{n\zeta_n}{1+\zeta_n \log^2 n}\right)^{\frac{1}{2s+d-1}}$. Specifying that $\lambda^* = (p_0^*, p_1^*, \dots, p_1^*, p_2^*, \dots)$ and $\widehat{T}_n = A/n$, Theorem 3.2 becomes for a graphon with regularity $s > 0$

$$\mathbb{E}\left[\delta_2^2\left(\lambda^*, \frac{\lambda(\widehat{T}_n)}{\zeta_n}\right)\right] = \mathcal{O}\left(\left(\frac{n\zeta_n}{1+\zeta_n \log^2 n}\right)^{\frac{-2s}{2s+d-1}}\right).$$

Figure 3.5 illustrates the estimation of the latitude and the envelope functions in some relatively sparse regimes.

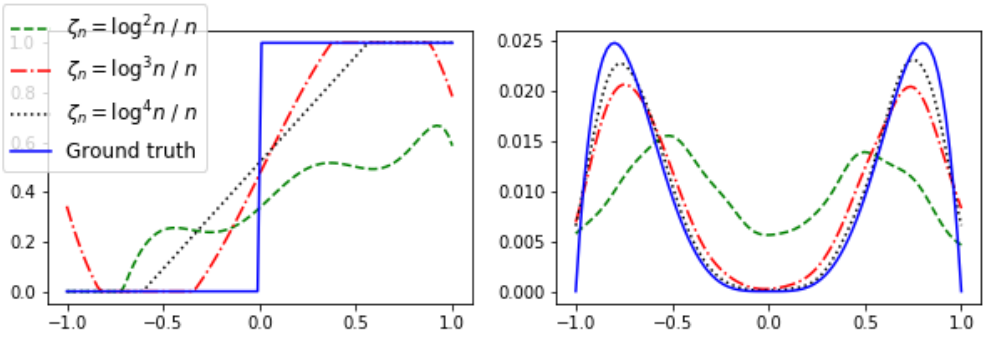


Figure 3.5: Results of our algorithms for graph of size 2000 with functions $\mathbf{p}^{(1)}$ and $f_{\mathcal{L}}^{(1)}$ of Eq.(3.12) and sparsity parameter $\zeta_n = \log^k n/n$, $k \in \{2, 3, 4\}$.

3.6 Experiments

In the following, we test our methods using different envelope and latitude functions. Note that a common choice of connection functions in RGGs are the *Rayleigh fading* activation functions which take the form

$$\mathcal{R}_{\zeta,\eta,r}(\rho) = \exp\left[-\zeta\rho^{\eta}\right], \quad \zeta > 0, \eta > 0.$$

Any Rayleigh function $\mathcal{R}_{\zeta,\eta}$ corresponds to the following envelope function

$$\mathbf{p}_{\zeta,\eta} : t \mapsto \mathcal{R}_{\zeta,\eta}(2(1-t)),$$

so that it holds

$$\forall x, y \in \mathbb{S}^{d-1}, \quad \mathbf{p}_{\zeta,\eta}(\langle x, y\rangle) = \mathcal{R}_{\zeta,\eta}(\|x - y\|_2).$$

Let us also denote for any $\alpha, \beta > 0$ $g(\cdot; \alpha, \beta)$ the density of the beta distribution $\mathcal{B}(\alpha, \beta)$ with parameters (α, β). In this paper, we will study the numerical results of our methods considering the following

envelope and latitude functions

$$\mathbf{p}^{(1)} : x \mapsto \mathbb{1}_{x\geq 0}, \qquad \mathbf{p}^{(2)} \equiv \mathbf{p}_{0.5,1}$$
$$f_{\mathcal{L}}^{(1)} : r \mapsto \begin{cases} \frac{1}{2}g(1-r;2,2) & \text{if } r \geq 0 \\ \frac{1}{2}g(1+r;2,2) & \text{otherwise} \end{cases}, \qquad f_{\mathcal{L}}^{(2)} : r \mapsto \frac{1}{2}g\left(\frac{1-r}{2};1,3\right)$$

$$\text{and} \quad \mathbf{p}^{(3)} \equiv \mathbf{p}_{0.25,3}$$
$$f_{\mathcal{L}}^{(3)} : r \mapsto \frac{1}{2}g\left(\frac{1-r}{2};2,2\right). \tag{3.12}$$

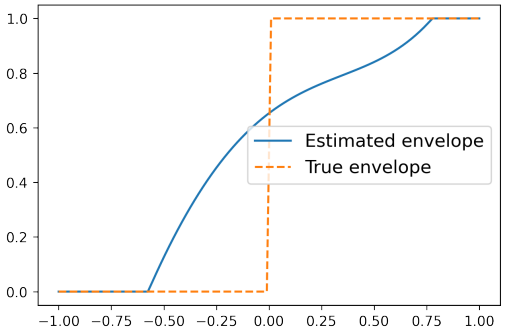
Note that considering the latitude function $f_{\mathcal{L}}^{(2)}$ (resp. $f_{\mathcal{L}}^{(3)}$) is equivalent to consider that one fourth of the Euclidean distance between consecutive latent positions is distributed as $Z \sim \mathcal{B}(1,3)$ (resp. $Z \sim \mathcal{B}(2,2)$). With Figures 3.6, 3.7 and 3.8, we present the results of our experiments for the three different settings described in Eq.(3.12). In each case, we work with a latent dimension $d = 4$ and we show:

1. the estimates of the envelope and latitude functions obtained with our adaptive procedure working the graph of 1500 nodes (see Figures (a) and (b)).

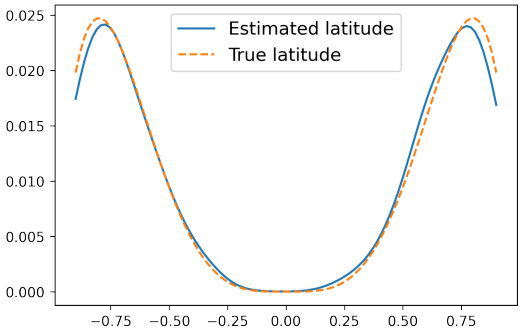
2. the corresponding clustering obtained by the SCCHEi algorithm for the resolution level R determined by the slope heuristic (see Figures (c)).

   Blue crosses represent the $\widetilde{R}$ eigenvalues of $\widehat{T}_n$ with the largest magnitude, which are used to form clusters corresponding to the $R+1$-first spherical harmonic spaces. The red plus are the estimated eigenvalues $(\hat{p}_k)_{0\leq k\leq R}$ (plotted with multiplicity) defined from the clustering given by our algorithm SCCHEi (see Eq. (3.7)). Those results show that SCCHEi achieves a relevant clustering of the eigenvalues of $\widehat{T}_n$ which allows us to recover the envelope function.

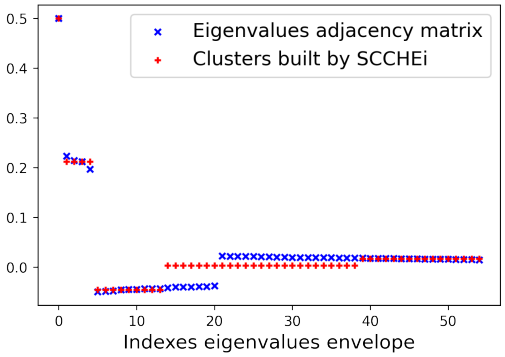
3. the errors between the estimated functions and the true ones in δ_2 metric and in L^2 norm for different size of graphs (see Figures (d) and (e)). We notice that a significant decrease of the δ_2 distance between spectra does not necessarily means that the L^2 norm between the estimated and the true envelope functions shrinks seriously. We refer in particular to Figures 3.6 and 3.8. The identifiability issue highlighted in Section 3.3.3 is one of the possible explanations of this phenomenon. Nevertheless, these experiments show that both the δ_2 and L^2 errors on our estimate of the envelope or the latitude functions are decreasing as the size of the graph is getting larger. Let us also recall that Theorem 3.6 ensures that the L^2 error on our estimate of the envelope function goes to zero as n grows when $\mathbf{p}$ has a finite number of non zeros eigenvalues that are all distinct.



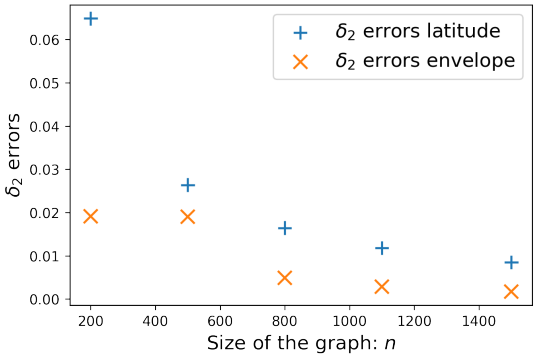
(a) Envelope function



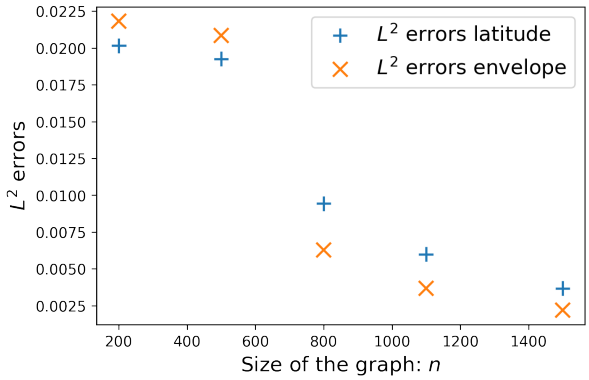
(b) Latitude function



(c) Eigenvalues envelope

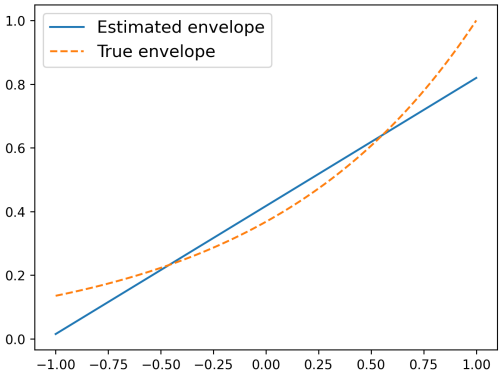


(d) δ_2 errors

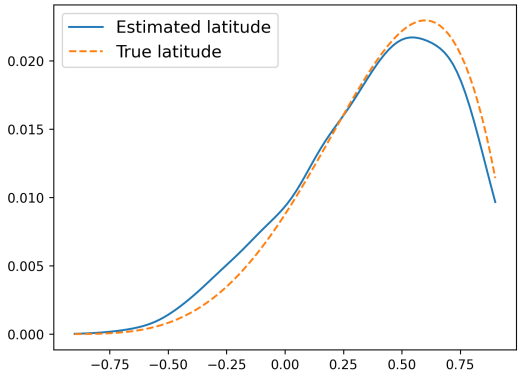


(e) L^2 errors

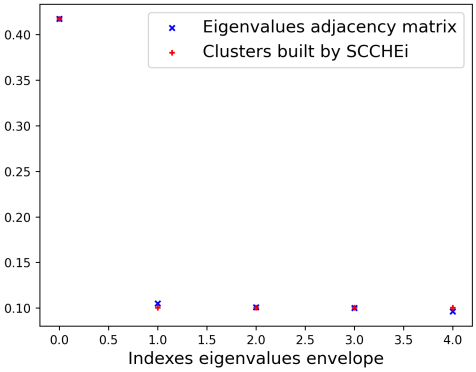
Figure 3.6: Results for $d = 4$, the envelope $\mathbf{p}^{(1)}$ and the latitude $f_{\mathcal{L}}^{(1)}$ of Eq.(3.12).



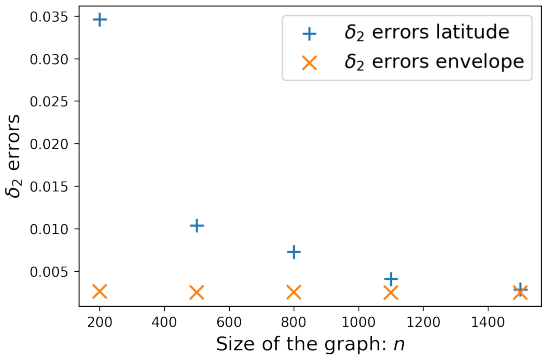
(a) Envelope function



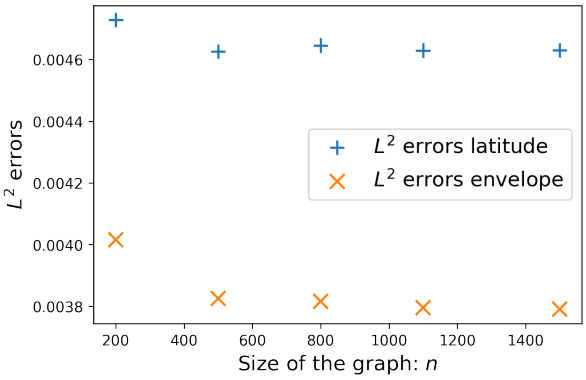
(b) Latitude function



(c) Eigenvalues envelope

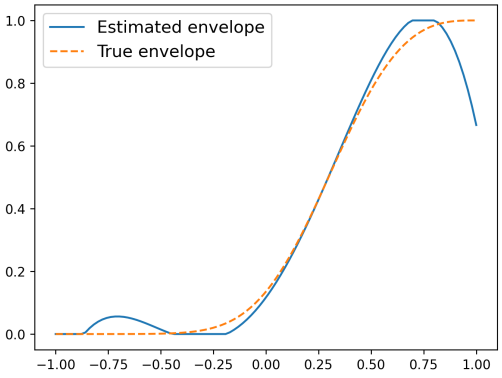


(d) δ_2 errors

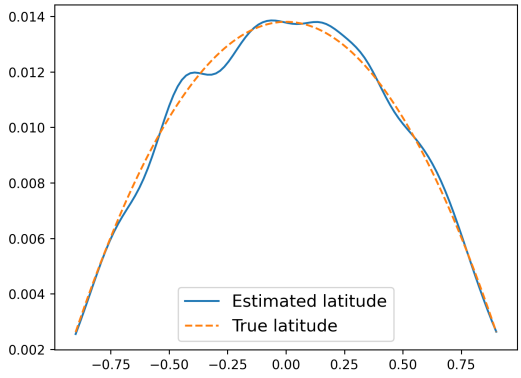


(e) L^2 errors

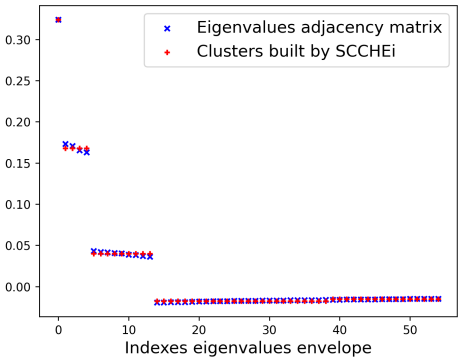
Figure 3.7: Results for $d = 4$, the envelope $\mathbf{p}^{(2)}$ and the latitude $f_{\mathcal{L}}^{(2)}$ of Eq.(3.12).



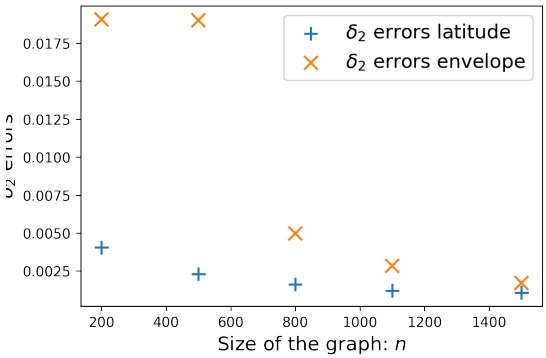
(a) Envelope function



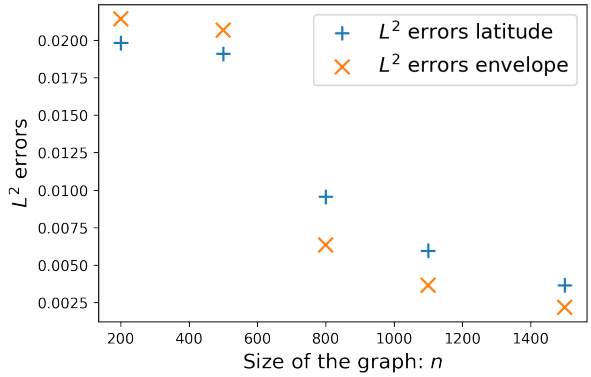
(b) Latitude function



(c) Eigenvalues envelope



(d) δ_2 errors



(e) L^2 errors

Figure 3.8: Results for $d = 4$, the envelope $\mathbf{p}^{(3)}$ and the latitude $f_{\mathcal{L}}^{(3)}$ of Eq.(3.12).

3.7 Applications

In this section, we apply the MRGG model to link prediction and hypothesis testing in order to demonstrate the usefulness of our approach as well as the estimation procedure.

3.7.1 Markovian Dynamic Testing

As a first application of our model, we propose a hypothesis test to statistically distinguish between an independent sampling the latent positions and a Markovian dynamic. The null is then set to $\mathbb{H}_0$: *nodes are independent and uniformly distributed on the sphere* (i.e., *no Markovian dynamic*). Our test is based on estimate $\hat{f}_{\mathcal{L}}$ of latitude and thus the null can be rephrased as $\mathbb{H}_0 : \ f_{\mathcal{L}} = f^0_{\mathcal{L}}$ where $f^0_{\mathcal{L}}$ is the latitude of uniform law, dynamic is then i.i.d. dynamic.

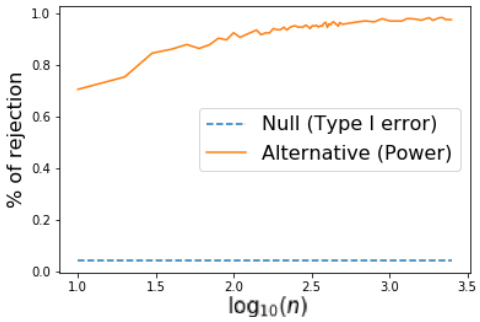


Figure 3.9: Hypothesis testing.

Figure 3.9 shows the power of a hypothesis test with level 5% (Type I error). One can use any *black-box goodness-of-fit test* comparing $\hat{f}_{\mathcal{L}}$ to $f^0_{\mathcal{L}}$, and we choose χ^2-test discretizing $(-1, 1)$ in 70 regular intervals. Rejection region is calibrated (i.e., threshold of the χ^2-test here) by *Monte Carlo simulations under the null*. It allows us to control Type I error as depicted by dotted blue line. We choose alternative given by Heaviside envelope $\mathbf{p}^{(1)}$ and latitude $f^{(1)}_{\mathcal{L}}$ of Eq.(3.12). We run our algorithm to estimate latitude from which we sample a batch to compute the χ^2-test statistic. We see that for graphs of size larger than $1,000$, the rejection rate is almost 1 under the alternative (Type II error is almost zero), the test is very powerful.

3.7.2 Link Prediction

Suppose that we observe a graph with n nodes. Link prediction is the task that consists in estimating the probability of connection between a given node of the graph and the upcoming node.

3.7.2.1 Bayes Link Prediction

We propose to show the usefulness of our model solving a link prediction problem. Let us recall that we do not estimate the latent positions but only the *pairwise distances* (embedding task is not necessary for our purpose). Denoting by $\mathrm{proj}_{X_n^\perp}(\cdot)$ the orthogonal projection onto the orthogonal complement of $\mathrm{Span}(\mathrm{X_n})$, the decomposition of $\langle X_i, X_{n+1}\rangle$ defined by

$$\begin{aligned}&\langle X_i, X_n\rangle\langle X_n, X_{n+1}\rangle \\ &+ \sqrt{1-\langle X_n, X_{n+1}\rangle^2}\sqrt{1-\langle X_i, X_n\rangle^2}\langle \frac{\mathrm{proj}_{X_n^\perp}(X_i)}{\|\mathrm{proj}_{X_n^\perp}(X_i)\|_2}, Y_{n+1}\rangle,\end{aligned} \tag{3.13}$$

shows that latent distances are enough for link prediction. Indeed, it can be achieved using a *forward step* on our Markovian dynamic, giving the posterior probability (cf. Definition 3.9) $\eta_i(\mathbf{D}_{1:n})$ defined by

$$\int_{[-1,1]^2} \mathbf{p}\left(\langle X_i, X_n\rangle r + \sqrt{1-r^2}\sqrt{1-\langle X_i, X_n\rangle^2}u\right) f_{\mathcal{L}}(r) w_{\frac{d-3}{2}}(u) \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d-2}{2})\sqrt{\pi}} dr du, \tag{3.14}$$

where $w_{\frac{d-3}{2}}(u) := (1-u^2)^{\frac{d-3}{2}-\frac{1}{2}}$ and where $\Gamma : a \in]0, +\infty[\mapsto \int_0^{+\infty} t^{a-1}e^{-t}dt$.

Definition 3.9. (Posterior probability function)
The posterior probability function η is defined for any latent pairwise distances $\mathbf{D}_{1:n} = (\langle X_i, X_j\rangle)_{1\le i,j\le n} \in$

$[-1,1]^{n\times n}$ by

$$\forall i \in [n], \quad \eta_i(\mathbf{D}_{1:n}) = \mathbb{P}\left(A_{i,n+1} = 1 \mid \mathbf{D}_{1:n}\right),$$

where $A_{i,n+1} \sim \text{Ber}\left(\mathbf{p}(\langle X_i, X_{n+1}\rangle)\right)$ is a random variable that equals 1 if there is an edge between nodes i and $n+1$, and is zero otherwise.

We consider a classifier g (cf. Definition 3.10) and an algorithm that, given some latent pairwise distances $\mathbf{D}_{1:n}$, estimates $A_{i,n+1}$ by putting an edge between nodes X_i and X_{n+1} if $g_i(\mathbf{D}_{1:n})$ is 1.

Definition 3.10. A *classifier* is a function which associates to any pairwise distances $\mathbf{D}_{1:n} = (\langle X_i, X_j\rangle)_{1\le i,j\le n}$, a label $(g_i(\mathbf{D}_{1:n}))_{i\in[n]} \in \{0,1\}^n$.

The risk of this algorithm is as in *binary classification*,

$$\begin{aligned}\mathcal{R}(g, \mathbf{D}_{1:n}) &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(g_i(\mathbf{D}_{1:n}) \neq A_{i,n+1} \mid \mathbf{D}_{1:n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\{(1-\eta_i(\mathbf{D}_{1:n}))\mathbb{1}_{g_i(\mathbf{D}_{1:n})=1} + \eta_i(\mathbf{D}_{1:n})\mathbb{1}_{g_i(\mathbf{D}_{1:n})=0}\right\}, \end{aligned} \tag{3.15}$$

where we used the independence between $A_{i,n+1}$ and $g_i(\mathbf{D}_{1:n})$ conditionally on $\varkappa(\mathbf{D}_{1:n})$. Pushing further this analogy, we can define the classification error of some classifier g by $L(g) = \mathbb{E}\left[\mathcal{R}(g, \mathbf{D}_{1:n})\right]$. Proposition 3.12 shows that the Bayes estimator - introduced in Definition 3.11 - is optimal for the risk defined in Eq.(3.15).

Definition 3.11. (Bayes estimator)
We keep the notations of Definition 3.9. The Bayes estimator g^* of $(A_{i,n+1})_{1\le i\le n}$ is defined by

$$\forall i \in [n], \quad g_i^*(\mathbf{D}_{1:n}) = \begin{cases} 1 & \text{if } \eta_i(\mathbf{D}_{1:n}) \ge \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 3.12. *(Optimality of the Bayes classifier for the risk $\mathcal{R}$)*
We keep the notations of Definitions 3.9 and 3.11. For any classifier g, it holds for all $i \in [n]$,

$$\begin{aligned}&\mathbb{P}\left(g_i(\mathbf{D}_{1:n}) \neq A_{i,n+1} \mid \mathbf{D}_{1:n}\right) - \mathbb{P}\left(g_i^*(\mathbf{D}_{1:n}) \neq A_{i,n+1} \mid \mathbf{D}_{1:n}\right) \\ &= 2\left|\eta_i(\mathbf{D}_{1:n}) - \frac{1}{2}\right| \times \mathbb{E}\left\{\mathbb{1}_{g_i(\mathbf{D}_{1:n}) \neq g_i^*(\mathbf{D}_{1:n})} \mid \mathbf{D}_{1:n}\right\},\end{aligned}$$

which immediately implies that

$$\mathcal{R}(g, \mathbf{D}_{1:n}) \ge \mathcal{R}(g^*, \mathbf{D}_{1:n}) \text{ and therefore } L(g) \ge L(g^*).$$

3.7.2.2 Heuristic for Link Prediction

One natural method to approximate the Bayes classifier from the previous section is to use the *plug-in approach*. This leads to the MRGG classifier introduced in Definition 3.13.

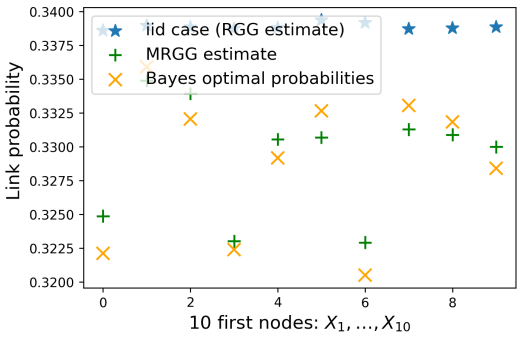
Definition 3.13. (The MRGG classifier)
For any n and any $i \in [n]$, we define $\hat{\eta}_i(\mathbf{D}_{1:n})$ as

$$\int \widehat{\mathbf{p}}\left(\widehat{r}_{i,n} r + \sqrt{1-r^2}\sqrt{1-\widehat{r}_{i,n}^2}\,u\right) \hat{f}_{\mathcal{L}}(r) w_{\frac{d-3}{2}}(u) \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d-2}{2})\sqrt{\pi}} dr du, \tag{3.16}$$

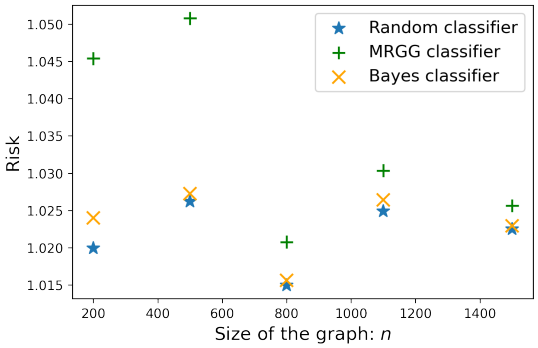
where $\widehat{\mathbf{p}}$ and $\hat{f}_{\mathcal{L}}$ denote respectively the estimate of the envelope function and the latitude function with our method and where $\widehat{r} := n\widehat{G}$. The MRGG classifier is defined by

$$\forall i \in [n], \quad g_i^{MRGG}(\mathbf{D}_{1:n}) = \begin{cases} 1 & \text{if } \hat{\eta}_i(\mathbf{D}_{1:n}) \ge \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

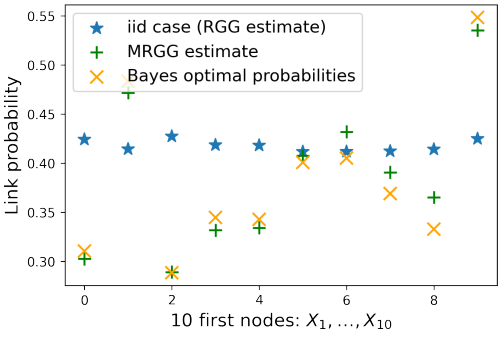
To illustrate our approach we work with a graph of 1500 nodes with $d = 4$, and we consider the envelope and latitude functions defined in Eq.(3.12). The plots on the left column of Figure 3.10 show that we are



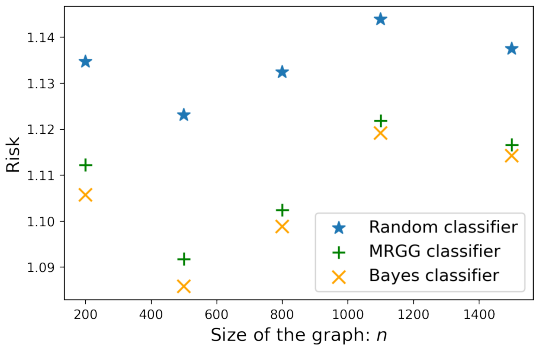
(a) Envelope $\mathbf{p}^{(1)}$, Latitude $f_{\mathcal{L}}^{(1)}$



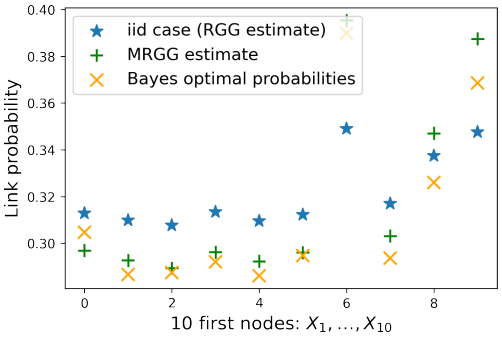
(b) Envelope $\mathbf{p}^{(1)}$, Latitude $f_{\mathcal{L}}^{(1)}$



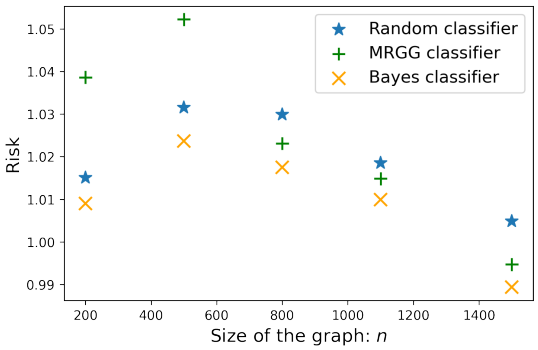
(c) Envelope $\mathbf{p}^{(2)}$, Latitude $f_{\mathcal{L}}^{(2)}$



(d) Envelope $\mathbf{p}^{(2)}$, Latitude $f_{\mathcal{L}}^{(2)}$



(e) Envelope $\mathbf{p}^{(3)}$, Latitude $f_{\mathcal{L}}^{(3)}$



(f) Envelope $\mathbf{p}^{(3)}$, Latitude $f_{\mathcal{L}}^{(3)}$

Figure 3.10: ← **On the left:** Link predictions between the future node X_{n+1} and the 10 first nodes $X_1, \ldots, X_{10}$. → **On the right:** Comparison between the risk (defined in Eq.(3.15)) of the MRGG classifier, the random classifier and the risk of the optimal Bayes classifier.

able to recover the probabilities of connection of the nodes already present in the graph with the coming node X_{n+1}. Using the decomposition of $\langle X_i, X_{n+1}\rangle$ given by Eq.(3.13), orange crosses are computed using Eq.(3.14). Green plus are computed similarly replacing $\mathbf{p}$ and $f_{\mathcal{L}}$ by their estimations $\widehat{\mathbf{p}}$ and $\hat{f}_{\mathcal{L}}$ following Eq.(3.16). Blue stars are computed using Eq.(3.14) by replacing $f_{\mathcal{L}}$ by $\frac{w_\beta}{\|w_\beta\|_1}$ (with $\beta = \frac{d-2}{2}$) which implicitly supposes that the points are sampled uniformly on the sphere.
With the plots on the left column of Figure 3.10, we compare the risk of the *random* classifier - whose guess $g_i(\mathbf{D}_{1:n})$ is a Bernoulli random variable with parameter given by the ratio of edges compared to complete graph - with the risk of the MRGG classifier (cf. Definition 3.13). These figures show that for a small number of nodes, the risk estimate provided by the MRGG classifier can be significantly far from the one of the Bayes classifier. However, when the number of nodes is getting larger, the MRGG classifier gives similar results compared to the optimal Bayes classifier. This risk estimate can be significantly smaller than the one of the random classifier (see for example the plots corresponding to the envelope $\mathbf{p}^{(2)}$ and the latitude $f_{\mathcal{L}}^{(2)}$).

3.8 Discussion

In this section, we want to push the investigation of the performance of our estimation methods as far as possible. In Section 3.8.1 we study the robustness of our methods under model mispecification before inspecting the influence of the mixing time of the Markov chain $(X_i)_{i\geq 1}$ on the estimation error in Section 3.8.2.
On a more theoretical side, we show that replacing the use of the complete linkage by the Ward distance in the SCCHEi algorithm, Theorem 3.3 might not be true anymore. We conclude with some remarks and by highlighting future research directions.

3.8.1 Robustness to model mispecification

We consider a mixture model for the sampling scheme of the latent position. We fix some $\epsilon \in (0,1)$ and we draw X_1 randomly on the sphere. Then at time step $i \geq 2$, the point X_i is sampled as follows:

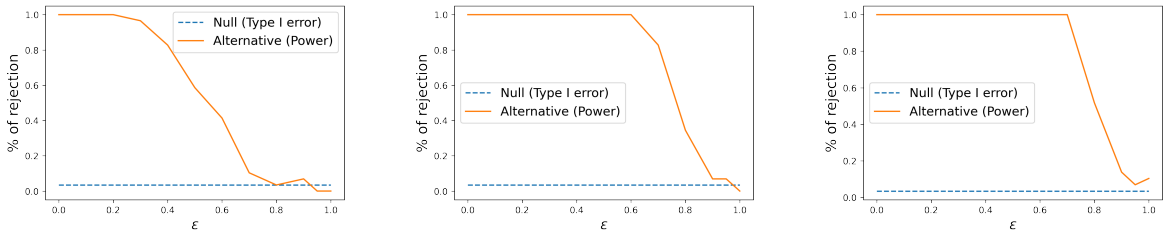
- with probability $1-\epsilon$, X_i is drawn following the Markovian dynamic described in Section 3.1 (based on X_{i-1}).
- with probability ϵ, X_i is drawn uniformly on the sphere.

Figure 3.11 and Figure 3.12 show the numerical results obtained under this mispecified model. We consider the hypothesis testing question presented in Section 3.7.1 with the same settings namely $d=3$ and the envelope and latitude functions $\mathbf{p}^{(1)}$ and $f_{\mathcal{L}}^{(1)}$ of Eq.(3.12). We can see that when $\epsilon = 0$, the power of our test is 1 and we always reject the null hypothesis (uniform sampling of the latent positions) under the alternative. On the contrary, when $\epsilon = 1$, the points are sampled uniformly on the sphere and we obtain a power of the order of the level of our test (i.e. 5%) as expected. The larger the sample size n is, the greater ϵ can be chosen while keeping a large power. In the case where $n = 1500$, one can afford to sample 75% of latent positions uniformly (and the rest using our Markovian sampling scheme) while keeping a power equal to 1. Figure 3.12 shows that the larger ϵ is, the closer the estimated latitude function is to $\frac{w_\beta}{\|w_\beta\|_1} \equiv \frac{1}{2}$ (since $d=3$) which corresponds to the density of a one-dimensional marginal of a uniform random point on $\mathbb{S}^{d-1}$.

3.8.2 Influence of mixing time on estimation error

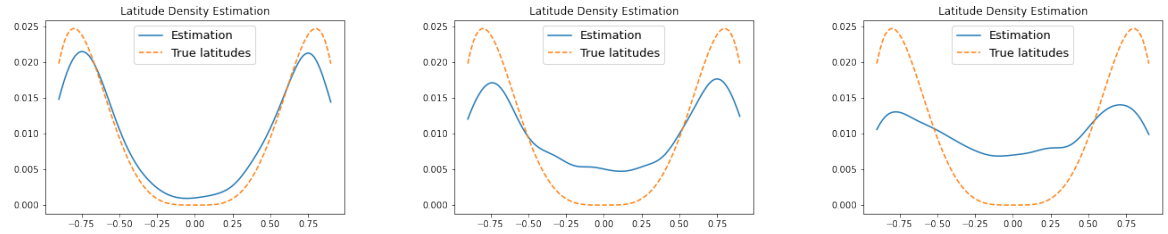
In order to assert that the dependence of the latent variables has an influence on the estimation of the unknown functions of our model, we would require a minimax bound. The derivation of such minimax result is still an open problem, even in the independent setting (cf. De Castro et al. [2019]). Nevertheless, by making explicit the constants involved in concentration inequalities, we can show that the mixing time of the latent Markovian dynamic affects our bound on the δ_2 error between spectra. For any $r^* \in (-1,1)$, let us consider the following latitude function

$$f_{\mathcal{L}}^{r^*}(r) := \frac{1}{I(r^*)}(1-r^2)^{\frac{d-3}{2}}\mathbb{1}_{r\in(r^*,1)}, \quad I(r^*) := \int_{r^*}^{1}(1-r^2)^{\frac{d-3}{2}}dr.$$



(a) $n = 200$ (b) $n = 500$ (c) $n = 1500$

Figure 3.11: Studying the robustness of our method under model mispecification. We study the evolution of power for Markovian Dynamic Testing when the mixture parameter ϵ ranges $(0, 1)$. We conduct this analysis for different values of n.



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.5$ (c) $\epsilon = 0.7$

Figure 3.12: Studying the robustness of the estimation of the latitude function under model mispecification. We plot our kernel density estimator of the latitude function for $n = 1500$, $d = 3$ and for $\epsilon \in \{0.1, 0.5, 0.7\}$. We use the envelope $\mathbf{p}^{(1)}$ and latitude function $f_{\mathcal{L}}^{(1)}$ defined in Eq.(3.12).

Note that the Markov transition kernel P of the chain $(X_i)_{i\geq 1}$ using this latitude function is the one that starting from a point $x \in \mathbb{S}^{d-1}$ samples uniformly a point in the set $\{z \in \mathbb{S}^{d-1} \mid \|x - z\|_2^2 \leq 2(1 - r^*)\}$. In particular, when $r^* = -1$, we recover the uniform distribution on the sphere. It is clear that the closer r^* to one, the larger the mixing time of the chain. One can show that for any $r^* \in (-1, 1)$, the chain is uniformly ergodic by proving that there exist an integer $m \geq 1$, a constant $\delta_m > 0$ and a probability measure ν such that

$$\forall x \in \mathbb{S}^{d-1},\ \forall A \in \Sigma, \quad P^m(x, A) \geq \delta_m \nu(A) \quad \text{(cf. Definition A.8)}. \tag{3.17}$$

Eq.(3.17) holds by considering for example $\nu = \pi$ the uniform distribution on the sphere. It is straightforward to show that the smallest integer $m(r^*) \geq 1$ satisfying Eq.(3.17) is larger than $\frac{2}{1-r^*}$.[2] Taking a closer look at the constants involved in the concentration inequality from Theorem 4.3 (cf. Chapter 4), we get that

$$\mathbb{E}\left[\delta_2^2(\lambda(\mathbb{T}_W), \lambda(T_n)) \vee \delta_2^2(\lambda(\mathbb{T}_W), \lambda^{R_{opt}}(\widehat{T}_n))\right] < C\left[\frac{n}{\log^2(n)}\right]^{-\frac{2s}{2s+d-1}},$$

where $C > m(r^*)^2\tau(r^*)^2\|f_{\mathcal{L}}^{r^*}\|_\infty$ and $\tau(r^*) \geq 1$ is the Orlicz norm of some regeneration time. Since for any $0 < r^* < 1$,

$$\begin{aligned} I(r^*) &= \int_{r^*}^{1} (1 - r^2)^{\frac{d-3}{2}} dr = \int_0^{1-r^*} e^{\frac{d-3}{2}\ln(1-(r+r^*)^2)} dr \\ &= (1 - (r^*)^2)^{\frac{d-3}{2}} \int_0^{1-r^*} e^{\frac{d-3}{2}\left\{\ln(1-(r+r^*)^2) - \ln(1-(r^*)^2)\right\}} dr \\ &\leq (1 - (r^*)^2)^{\frac{d-3}{2}} \int_0^{1-r^*} e^{-\frac{d-3}{2}\left\{\frac{2rr^* + r^2}{1-(r^*)^2}\right\}} dr \end{aligned}$$

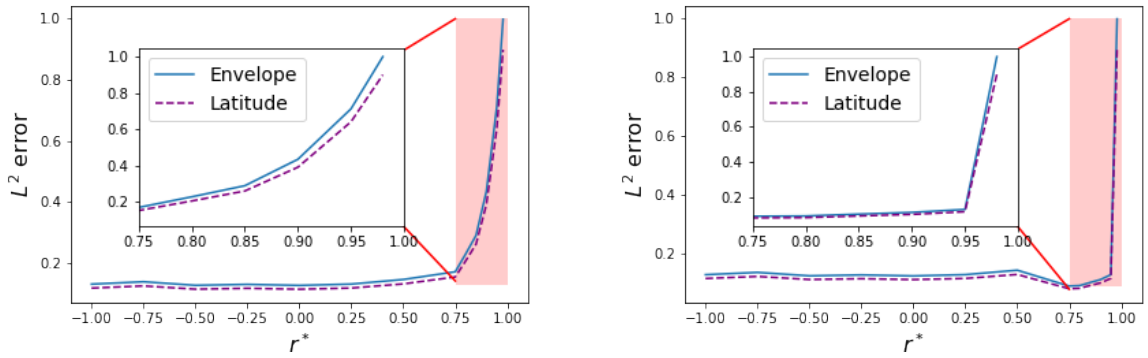
[2] Indeed, the latitude function $f_{\mathcal{L}}^{r^*}$ allows to make a jump at each time step of size at most $1 - r^*$. Since the length of the shortest arc on $\mathbb{S}^{d-1}$ joining the north pole to the south pole is 2, the result follows.

$$\leq (1-(r^*)^2)^{\frac{d-3}{2}} \int_0^{1-r^*} e^{-\frac{d-3}{2}\{2rr^*+r^2\}} dr$$
$$\leq (1-(r^*)^2)^{\frac{d-3}{2}} \int_0^{1} e^{-\frac{d-3}{2}\{2rr^*\}} dr$$
$$\leq (1-(r^*)^2)^{\frac{d-3}{2}} \left(1 \wedge \frac{1}{r^*(d-3)}\right),$$

we get that $\|f_{\mathcal{L}}^{r^*}\|_\infty \geq \frac{1}{I(r^*)}(1-(r^*)^2)^{\frac{d-3}{2}} \geq r^*(d-3)$. Finally we obtain

$$C > \frac{2r^*}{1-r^*}(d-3),$$

where $r^* \mapsto \frac{2r^*}{1-r^*}(d-3)$ is increasing in r^* and diverges to $+\infty$ when $r^* \to 1^-$. Hence, the closer r^* is to one, the slower the chain is mixing, and the poorer is our bound.
Figure 3.13 presents the result of the simulations using the latitude function $f_{\mathcal{L}}^{r^*}$ and the envelope function $\mathbf{p} : t \mapsto \mathbb{1}_{t\geq 0}$. We compute the L^2 error between the true and the estimated envelope functions (respectively the true and the estimated latitude functions). When r^* is getting closer to 1, the chain is mixing slowly and we need to increase the sample size if we want to prevent the L^2 errors from blowing up. Graphs have been generated with a latent dimension $d = 3$ and by sampling the latent positions using our isotropic sampling procedure with latitude function $f_{\mathcal{L}}^{r^*}$.



(a) $n = 200$ (b) $n = 1500$

Figure 3.13: Studying the influence of the mixing time of the chain on the L^2 errors between (i) the envelope function and its estimate (using our adaptive procedure), and (ii) the latitude function and its estimate obtained with a kernel estimator.

3.8.3 Choice of the clustering algorithm for the SCCHEi

The SCCHEi algorithm relies on the clustering of the eigenvalues of the adjacency matrix provided by the HAC with complete linkage. In this section, we motivate the use of the HAC algorithm with complete linkage by showing that the theoretical results from Section 3.3.3 could be much more involved to establish by using another clustering procedure. Indeed, if one would consider for example the HAC with the Ward distance, the theoretical result obtained for the correctness of the SCCHEi algorithm (cf. Theorems 3.3 and 3.6) is likely to be no longer true (even if the sample size n is chosen arbitrarily large). Let us show this on a simple example.

We fix a resolution level $R = 2$ and we consider some $\Delta^G > 0$. We set $p_0^* = 4\Delta^G$, $p_1^* = 3\Delta^G$, $p_2^* = 2\Delta^G$, and $p_k^* = 0$ for all $k \geq 3$. Let us consider some $g \in (0, \Delta^G/4)$ that can be taken arbitrarily small. Let us denote $\lambda^R(\widehat{T}_n) = (\hat{\lambda}_1, \dots, \hat{\lambda}_{\widetilde{R}}, 0, 0 \dots)$ and assume that it holds $\hat{\lambda}_1 = p_0^*$, $\hat{\lambda}_2 = \cdots = \hat{\lambda}_{d+1} = p_1^*$ (we recall that $d_1 = d$), $\hat{\lambda}_{d+2} = \cdots = \hat{\lambda}_{d+1+\lfloor d_2/2 \rfloor} = p_2^* + g$ and $\hat{\lambda}_{d+2+\lfloor d_2/2 \rfloor} = \cdots = \hat{\lambda}_{1+d+d_2} = p_2^* - g$. To simplify the presentation, we will assume in the following that $d_2 = \frac{(d+1)d}{2} - 1$ is even (which holds for example if $d = 2k$ for any $k \geq 1$ odd). Figure 3.14 gives a visualization of this example.

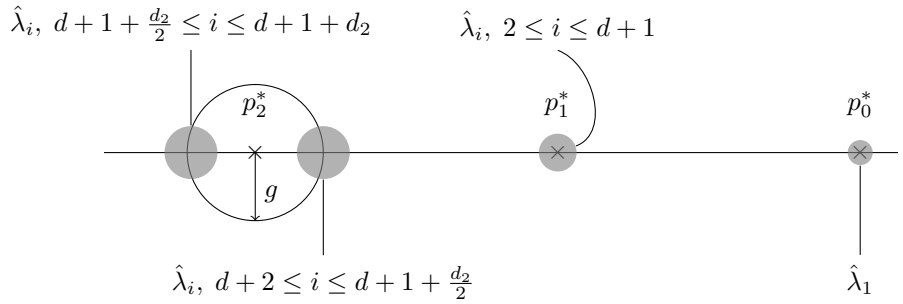


Figure 3.14: Visualization of the eigenvalues of the envelope function of our example.

Applying the HAC algorithm (with the Ward distance) to the eigenvalues $(\hat{\lambda}_1, \dots, \hat{\lambda}_{\widetilde{R}})$, it is obvious that the state reached after $\widetilde{R} - 4 = 1 + d + d_2 - 4$ iterations in the HAC procedure will be

$$\begin{aligned}
\widehat{\mathcal{G}}_0 :=& \{\hat{\lambda}_1\} \\
\widehat{\mathcal{G}}_1 :=& \{\hat{\lambda}_i \mid 2 \leq i \leq d\} \\
\widehat{\mathcal{G}}_2 :=& \{\hat{\lambda}_i \mid d+2 \leq i \leq d+1+d_2/2\} \\
\widehat{\mathcal{G}}_3 :=& \{\hat{\lambda}_i \mid d+2+d_2/2 \leq i \leq 1+d+d_2\}
\end{aligned}$$

Hence, in order to understand which clusters will be merged at the next step of the HAC algorithm, we compute the Ward distance between the different clusters.
Let us recall that for two finite and non-empty sets $S, S' \subset \mathbb{R}$ with respective cardinality $|S|$ and $|S'|$, the Ward distance between S and S' is given by

$$d_W(S, S') := \frac{|S| \times |S'|}{|S| + |S'|} \left(\frac{1}{|S|} \sum_{x_s \in S} x_s - \frac{1}{|S'|} \sum_{x'_s \in S'} x'_s \right)^2 .$$

Ward distances between clusters

	$\widehat{\mathcal{G}}_1$	$\widehat{\mathcal{G}}_2$	$\widehat{\mathcal{G}}_3$
$\widehat{\mathcal{G}}_0$	$\frac{d}{d+1}(\Delta^G)^2$	$\frac{d_2}{d_2+2}(2\Delta^G - g)^2$	$\frac{d_2}{d_2+2}(2\Delta^G + g)^2$
$\widehat{\mathcal{G}}_1$		$\frac{d \times d_2}{2d+d_2}(\Delta^G - g)^2$	$\frac{d \times d_2}{2d+d_2}(\Delta^G + g)^2$
$\widehat{\mathcal{G}}_2$			$d_2 \times g^2$

We deduce that all Ward distances between pair of clusters are scaling at least linearly with d except the Ward distances between $\widehat{\mathcal{G}}_0$ and the other three clusters $\widehat{\mathcal{G}}_1$, $\widehat{\mathcal{G}}_2$ and $\widehat{\mathcal{G}}_3$. Indeed, for any $i \in \{1, 2, 3\}$, $d_W(\widehat{\mathcal{G}}_0, \widehat{\mathcal{G}}_i)$ remains bounded independently of the latent dimension d. Hence, for any $g \in (0, \Delta^G/4)$ which can be chosen arbitrarily small, one can take d large enough to ensure that

$$\max \left\{ d_W(\widehat{\mathcal{G}}_0, \widehat{\mathcal{G}}_i) \ , \ i \in \{1, 2, 3\} \right\} < d_W(\widehat{\mathcal{G}}_2, \widehat{\mathcal{G}}_3). \tag{3.18}$$

We deduce that for any $g \in (0, \Delta^G/4)$, we can choose d large enough to ensure that Eq.(3.18) holds and thus the clusters merged between depths 4 and 3 from the root of the HAC's tree will not be $\widehat{\mathcal{G}}_2$ and $\widehat{\mathcal{G}}_3$. This means that the state obtained at depth 3 from the root is not of type $(\mathcal{S})$ (in the sense defined in Lemma 3.5).
If this is not a sufficient condition to state that the SCCHEi will fail to recover the correct clusters, this example shows that the use of Ward distance can lead to some unexpected clustering of the eigenvalues. Our example proves that using the HAC algorithm with the Ward distance, the result of Lemma 3.5 does not hold anymore. Namely, regardless of how large the sample size is chosen, there are situations (in particular for a large latent dimension) where the states of type $(\mathcal{S})$ (cf. Lemma 3.5) are never reached

in the HAC tree with the Ward distance. Hence obtaining a theoretical guarantee for the clustering provided by the SCCHEi in this framework may be impossible or at least much more involved.

3.8.4 Concluding remarks

3.8.4.1 Estimation of the latent dimension

The proposed methods implicitly assume that the latent dimension d is known. Araya and De Castro [2019] proved that the latent dimension d can be easily recovered in practice for n large enough provided that the spectral gap condition (3.10) holds. In the following, we briefly describe their approach. Given some matrix $\widehat{T}_n$ as input and some set of candidates $\mathcal{D}$ for the dimension d (typically $\mathcal{D} = \{2, 3, \dots, d_{\max}\}$), apply the Algorithm HEiC (cf. Algorithm 3 in Section 3.4.3) for any $d_c \in \mathcal{D}$ and store the returned value $gap := gap(d_c)$. Let us recall that $gap(d_c)$ corresponds to the largest gap between a bulk of d_c eigenvalues of $\widehat{T}_n$ and the rest of the spectrum (see the definition of Gap_1 in Section 3.4.3 for details). Once we have computed the different gaps, we pick the candidate d_c that led to the largest one. Given the guarantees provided by Proposition 3.24, the previously described procedure will find the correct dimension, with high probability (on the event $\mathcal{E}$ with the notations of Proposition 3.24), if the true dimension of the latent space is in the candidate set $\mathcal{D}$.

3.8.4.2 Future research directions

Our work encourages the development of growth model in random graphs and in particular the derivation of similar results in MRGGs with other latent spaces. It would be also desirable to extend our methods to the case where we consider more complex Markovian sampling of the latent positions, typically one that is not isotropic. Our work leaves open the question of getting a theoretical guarantee for the estimation of the latitude function. If we proved (with Theorem 3.8) that we can consistently estimate the Gram matrix of the latent positions in Frobenius norm, this is not sufficient to ensure that our kernel density estimator is consistent since we cannot ensure that $\frac{1}{n-1}\sqrt{\sum_{i=2}^{n}(r_i - \hat{r}_i)^2}$ tends to 0 as n goes to $+\infty$. Deriving a theoretical result regarding the estimation of the latitude function seems challenging and we believe that it would require significantly different proof techniques.

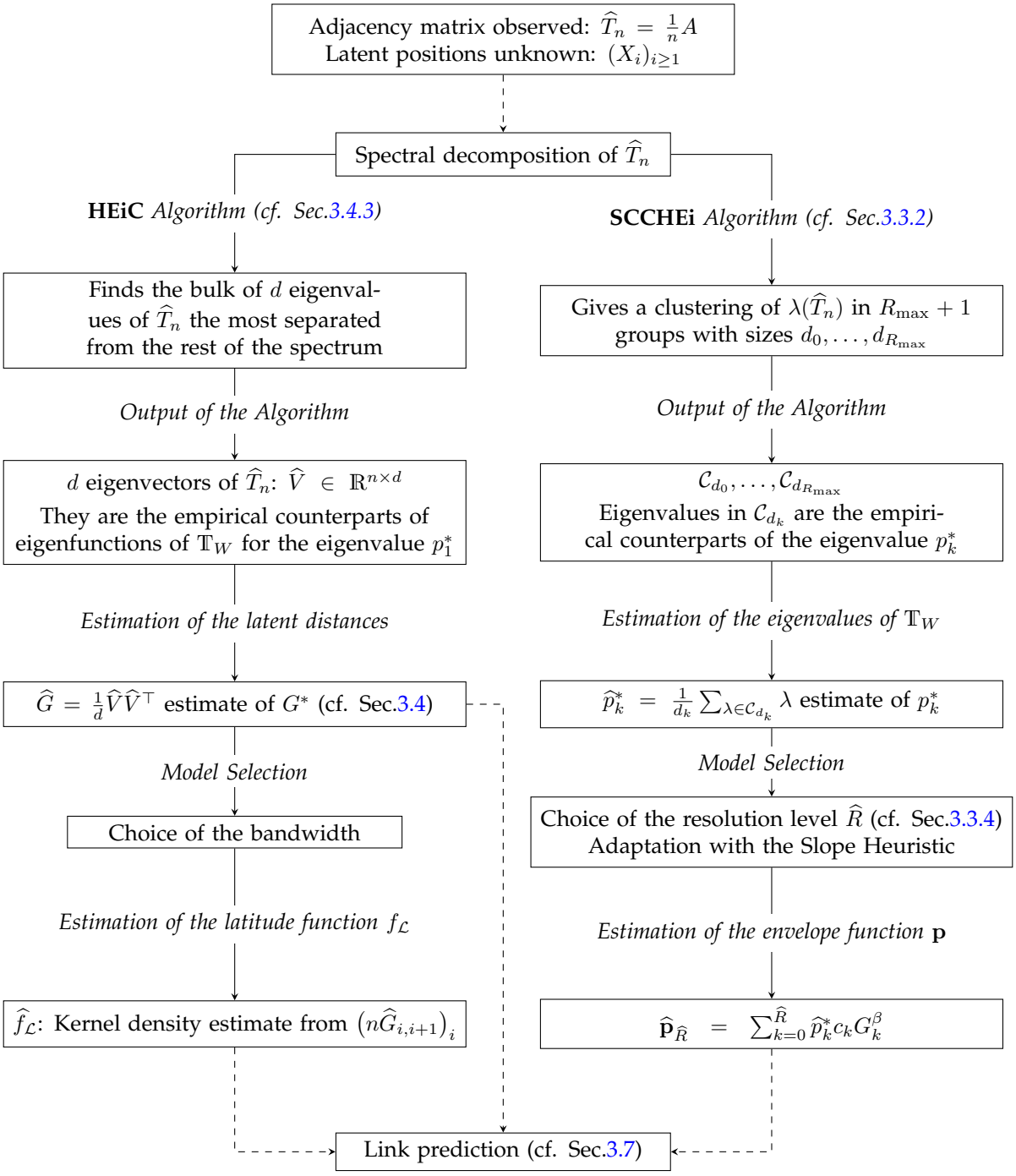


Figure 3.15: Synthetic presentation of the different estimation procedures.

3.9 Properties of the Markov chain

In the following, we denote $\lambda_{Leb} \equiv \lambda_{Leb,d}$ the Lebesgue measure on $\mathbb{S}^{d-1}$ and $\lambda_{Leb,d-1}$ the Lebesgue measure on $\mathbb{S}^{d-2}$. Using [Dai and Xu, 2013, Section 1.1], it holds $b_d := \int_{x\in\mathbb{S}^{d-1}} \lambda_{Leb,d}(dx) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$. Let P be the Markov operator of the Markov chain $(X_i)_{i\geq 1}$. By abuse of notation, we will also denote $P(x,\cdot)$ the density of the measure $P(x,dz)$ with respect to $\lambda_{Leb}(dz)$. For any $x,z\in\mathbb{S}^{d-1}$, we denote $R_x^z \in \mathbb{R}^{d\times d}$ a rotation matrix sending x to z (i.e. $R_x^z x = z$) and keeping $\mathrm{Span}(x,z)^\perp$ fixed. In the following, we denote $e_d := (0,0,\dots,0,1)\in\mathbb{R}^d$.

3.9.1 Invariant distribution and reversibility for the Markov chain

Reversibility of the Markov chain $(X_i)_{i\geq 1}$.

Lemma 3.14. *For all $x,z\in\mathbb{S}^{d-1}$, $P(x,z) = P(z,x) = P(e_d, R_z^{e_d}x)$.*

Proof of Lemma 3.14. Using our model described in Section 3.3, we get $X_2 = rX_1 + \sqrt{1-r^2}Y$ where conditionally on X_1, Y is uniformly sampled on $\mathcal{S}(X_1) := \{q\in\mathbb{S}^{d-1} \,:\, \langle q, X_1\rangle = 0\}$, and where r has density $f_{\mathcal{L}}$ on $[-1,1]$. Let us consider a Gaussian vector $W\sim\mathcal{N}(0,I_d)$. Using the Cochran's theorem and Lemma 3.15, we know that conditionally on X_1, the random variable $\frac{W-\langle W,X_1\rangle X_1}{\|W-\langle W,X_1\rangle X_1\|_2}$ is distributed uniformly on $\mathcal{S}(X_1)$.

Lemma 3.15. *Let $W\sim\mathcal{N}(0,I_d)$. Then, $\frac{W}{\|W\|_2}$ is distributed uniformly on the sphere $\mathbb{S}^{d-1}$.*

In the following, we denote $\overset{(d)}{=}$ the equality in distribution sense. We have conditionally on X_1

$$R_{X_1}^{e_d}\frac{W-\langle W,X_1\rangle X_1}{\|W-\langle W,X_1\rangle X_1\|_2} = \frac{\hat{W}-\langle\hat{W},e_d\rangle e_d}{\|\hat{W}-\langle\hat{W},e_d\rangle e_d\|_2},$$

where $\hat{W} = R_{X_1}^{e_d}W\sim\mathcal{N}(0,I_d)$. Using Cochran's theorem, we know that $\hat{W}-\langle\hat{W},e_d\rangle e_d$ is a centered normal vector with covariance matrix the orthographic projection matrix onto the space $\mathrm{Span}(e_d)^\perp$, leading to

$$\hat{W}-\langle\hat{W},e_d\rangle e_d \overset{(d)}{=} \begin{bmatrix} Y \\ 0 \end{bmatrix},$$

where $Y\sim\mathcal{N}(0,I_{d-1})$. Using Lemma 3.15, we conclude that conditionally on X_1, the random variable $\frac{W-\langle W,X_1\rangle X_1}{\|W-\langle W,X_1\rangle X_1\|_2}$ is distributed uniformly on $\mathcal{S}(X_1)$ (because the distribution of Y is invariant by rotation).

We deduce that

$$\begin{aligned} X_2 &\overset{(d)}{=} rX_1 + \sqrt{1-r^2}\frac{W-\langle W,X_1\rangle X_1}{\|W-\langle W,X_1\rangle X_1\|_2} \\ &\overset{(d)}{=} rX_1 + \sqrt{1-r^2}\frac{R_{X_2}^{X_1}W' - \langle R_{X_2}^{X_1}W', X_1\rangle X_1}{\|R_{X_2}^{X_1}W' - \langle R_{X_2}^{X_1}W', X_1\rangle X_1\|_2}, \end{aligned}$$

where $W' := R_{X_1}^{X_2}W$. Note that $W'\in\mathbb{R}^d$ is also a standard centered Gaussian vector because this distribution is invariant by rotation. Since $\langle R_{X_2}^{X_1}W', X_1\rangle = \langle W', X_2\rangle$ and $\|R_{X_2}^{X_1}q\|_2 = \|q\|_2,\ \forall q\in\mathbb{S}^{d-1}$, we deduce that

$$X_2 - rX_1 \overset{(d)}{=} R_{X_2}^{X_1}\left[\sqrt{1-r^2}\frac{W'-\langle W',X_2\rangle X_2}{\|W'-\langle W',X_2\rangle X_2\|_2}\right]. \tag{3.19}$$

$R_{X_1}^{X_2}$ is the rotation that sends X_1 to X_2 keeping the other dimensions fixed. Let us denote $a_1 := X_1$, $a_2 := \frac{X_2-rX_1}{\|X_2-rX_1\|_2}$ and complete the linearly independent family (a_1,a_2) in an orthonormal basis of $\mathbb{R}^d$

given by $a := (a_1, a_2, \dots, a_d)$. Then, the matrix of $R_{X_1}^{X_2}$ in the basis a is

$$\begin{bmatrix} r & -\sqrt{1-r^2} & 0_{d-2}^\top \\ \sqrt{1-r^2} & r & 0_{d-2}^\top \\ 0_{d-2} & 0_{d-2} & I_{d-2} \end{bmatrix}.$$

We deduce that

$$\begin{aligned}
\left(R_{X_2}^{X_1}\right)^{-1}(X_2 - rX_1) &= R_{X_1}^{X_2}(X_2 - rX_1) \\
&= \|X_2 - rX_1\|_2 R_{X_1}^{X_2}\left(\frac{X_2 - rX_1}{\|X_2 - rX_1\|_2}\right) \\
&= \|X_2 - rX_1\|_2 R_{X_1}^{X_2} a_2 \\
&= \|X_2 - rX_1\|_2 \left[-\sqrt{1-r^2}a_1 + ra_2\right] \\
&= -\sqrt{1-r^2}\|X_2 - rX_1\|_2 X_1 + rX_2 - r^2 X_1 \\
&= -(1-r^2)X_1 + rX_2 - r^2 X_1 \\
&= -X_1 + rX_2.
\end{aligned}$$

Going back to Eq.(3.19), we deduce that

$$X_1 \overset{(d)}{=} rX_2 + \sqrt{1-r^2}\frac{\tilde{W} - \langle \tilde{W}, X_2\rangle X_2}{\|\tilde{W} - \langle \tilde{W}, X_2\rangle X_2\|_2}, \tag{3.20}$$

where $\tilde{W} = -W'$ is also a standard centered Gaussian vector in $\mathbb{R}^d$. Thus, we proved the first equality of Lemma 3.14. Based on Eq.(3.20) we have,

$$\begin{aligned}
R_{X_2}^{e_d} X_1 &\overset{(d)}{=} rR_{X_2}^{e_d} X_2 + \sqrt{1-r^2}\frac{R_{X_2}^{e_d}\tilde{W} - \langle \tilde{W}, X_2\rangle R_{X_2}^{e_d} X_2}{\|\tilde{W} - \langle \tilde{W}, X_2\rangle X_2\|_2} \\
&= re_d + \sqrt{1-r^2}\frac{R_{X_2}^{e_d}\tilde{W} - \langle R_{X_2}^{e_d}\tilde{W}, e_d\rangle e_d}{\|R_{X_2}^{e_d}\tilde{W} - \langle R_{X_2}^{e_d}\tilde{W}, e_d\rangle e_d\|_2},
\end{aligned}$$

which proves that $P(e_d, R_{x_2}^{e_d} x_1) = P(x_2, x_1)$ for any $x_1, x_1 \in \mathbb{S}^{d-1}$ because $R_{X_2}^{e_d}\tilde{W}$ is again a standard centered Gaussian vector in $\mathbb{R}^d$. □

Stationary distribution of the Markov chain.

Proposition 3.16. *The uniform distribution on the sphere $\mathbb{S}^{d-1}$ is a stationary distribution of the Markov chain $(X_i)_{i\geq 1}$.*

Proof of Proposition 3.16. Let us consider $z \in \mathbb{S}^{d-1}$. We have using Lemma 3.14,

$$\int_{x\in\mathbb{S}^{d-1}} P(x,z)\lambda_{Leb}(dx) = \int_{x\in\mathbb{S}^{d-1}} P(z,x)\lambda_{Leb}(dx) = 1,$$

which proves that the uniform distribution on the sphere is a stationary distribution of the Markov chain. □

3.9.2 Ergodicity of the Markov chain

Our results hold under the condition that the Markov chain $(X_i)_{i\geq 1}$ is uniformly ergodic (cf. Assumption A). In this section, we provide a sufficient condition on the latitude function $f_{\mathcal{L}}$ for uniform ergodicity to hold.

Lemma 3.17. *We consider that $f_{\mathcal{L}}$ is bounded away from zero. Then, the Markov chain $(X_i)_{i\geq 1}$ is π-irreducible and aperiodic.*

Lemma 3.18. *We consider that $f_{\mathcal{L}}$ is bounded away from zero. Then the Markov chain $(X_i)_{i\geq 1}$ is uniformly ergodic.*

Proof of Lemmas 3.17 and 3.18. Considering for π the uniform distribution on $\mathbb{S}^{d-1}$, we get that for any $x \in \mathbb{S}^{d-1}$ and any $A \subset \mathbb{S}^{d-1}$ with $\pi(A) > 0$,

$$
\begin{aligned}
&P(x, A)\\
&= \int_{z\in A} P(x,z)\frac{\lambda_{Leb,d}(dz)}{b_d}\\
&= \int_{z\in A} P(e_d, R_x^{e_d} z)\frac{\lambda_{Leb,d}(dz)}{b_d} \quad \text{(Using Lemma 3.14)}\\
&= \int_{z\in R_x^{e_d} A} P(e_d, z)\frac{\lambda_{Leb,d}(dz)}{b_d}\\
&\text{(Using the change of variable } z \mapsto R_x^{e_d} z \text{ with } R_x^{e_d} A = \{R_x^{e_d} a : a \in A\})\\
&= \int_{r\in[-1,1]}\int_{\xi\in\mathbb{S}^{d-2}} f_{\mathcal{L}}(r) 1_{(\xi^\top\sqrt{1-r^2}, r)^\top \in R_x^{e_d} A} dr \frac{\lambda_{Leb,d-1}(d\xi)}{b_{d-1}b_d}\\
&\geq \inf_{s\in[-1,1]} f_{\mathcal{L}}(s) \int_{r\in[-1,1]}\int_{\xi\in\mathbb{S}^{d-2}} 1_{(\xi^\top\sqrt{1-r^2}, r)^\top \in R_x^{e_d} A} dr \frac{\lambda_{Leb,d-1}(d\xi)}{b_{d-1}b_d}\\
&\geq \inf_{s\in[-1,1]} f_{\mathcal{L}}(s) \int_{r\in[-1,1]}\int_{\xi\in\mathbb{S}^{d-2}} 1_{(\xi^\top\sqrt{1-r^2}, r)^\top \in R_x^{e_d} A} \left(1-r^2\right)^{\frac{d-3}{2}} \frac{dr\lambda_{Leb,d-1}(d\xi)}{b_{d-1}b_d}\\
&= \frac{1}{b_{d-1}} \inf_{s\in[-1,1]} f_{\mathcal{L}}(s)\pi(R_x^{e_d} A) = \frac{1}{b_{d-1}} \inf_{s\in[-1,1]} f_{\mathcal{L}}(s)\pi(A),
\end{aligned}
$$

since π is invariant by rotation and $f_{\mathcal{L}}$ is bounded away from zero. We also used that $\int_{-1}^{1}(1-r^2)^{\frac{d-3}{2}}dr = \frac{b_d}{b_{d-1}}$. This result means that the whole space $\mathbb{S}^{d-1}$ is a small set. Hence, the Markov chain is uniformly ergodic (cf. [Meyn and Tweedie, 1993, Theorem 16.0.2]) and thus aperiodic and π-irreducible. □

3.9.3 Computation of the absolute spectral gap of the Markov chain

Thanks to Proposition A.11 (in Appendix A), we know that if $f_{\mathcal{L}}$ is such that $(X_i)_{i\geq 1}$ is uniformly ergodic, the Markov chain has an absolute spectral gap (cf. Definition A.10). In the following, we show that this absolute spectral gap is equal to 1.
Keeping notations of Appendix A, let us consider $h \in L_0^2(\pi)$ such that $\|h\|_\pi = 1$. Then

$$
\begin{aligned}
\|Ph\|_\pi^2 &= \int_{x\in\mathbb{S}^{d-1}} \left(\int_{y\in\mathbb{S}^{d-1}} P(x,dy)h(y)\right)^2 \pi(dx)\\
&= \int_{x\in\mathbb{S}^{d-1}} \left(\int_{y\in\mathbb{S}^{d-1}} P(x,y)h(y)\pi(dy)\right)^2 \pi(dx)\\
&= \int_{x\in\mathbb{S}^{d-1}} \left(\int_{y\in\mathbb{S}^{d-1}} P(e_d, R_y^{e_d} x)h(y)\pi(dy)\right)^2 \pi(dx) \quad \text{(Using Lemma 3.14)}\\
&= \int_{x\in\mathbb{S}^{d-1}} \left(\int_{y\in\mathbb{S}^{d-1}} P(e_d, x)h(y)\pi(dy)\right)^2 \pi(dx)\\
&\qquad \text{(Using the rotational invariance of } \pi)\\
&= \int_{x\in\mathbb{S}^{d-1}} P(e_d,x)^2 \left(\int_{y\in\mathbb{S}^{d-1}} h(y)\pi(dy)\right)^2 \pi(dx)\\
&= 0,
\end{aligned}
$$

where the last equality comes from $h \in L_0^2(\pi)$. Hence, the Markov chain $(X_i)_{i\geq 1}$ has 1 for absolute spectral gap.

3.10 Proofs

3.10.1 Proofs of the two key lemmas for Theorem 3.3

In the proofs of Lemma 3.4 and Lemma 3.5 provided in this section, we keep the notations and the assumptions used in the proof of Theorem 3.3. To ease the reading of this section, we recall here important notations.
We denoted

$$\Delta^G = \min_{0\le k\neq l\le R,\ p_k^*\neq p_l^*} |p_k^* - p_l^*| \wedge \min_{0\le k\le R,\ p_k^*\neq 0} |p_k^*| > 0.$$

For any $g \in (0, \frac{\Delta^G}{4})$, the proof of Theorem 3.2 (cf. Section 3.10.3) ensures that for n large enough it holds

$$\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) \le g^2. \tag{3.21}$$

Let us finally recall (cf. Section 3.1) that

$$\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) = \inf_{\sigma\in\mathfrak{S}} \sum_{i\ge 1} \left(\lambda(\mathbb{T}_{W_R})_{\sigma(i)} - \lambda^R(\widehat{T}_n)_i\right)^2. \tag{3.22}$$

3.10.1.1 Proof of Lemma 3.4

We denote σ^* a permutation achieving the minimum in Eq.(3.22).
• First we show that we can choose σ^* such that $\sigma^*(\{1,\dots,\widetilde{R}\}) = \{1,\dots,\widetilde{R}\}$. We recall that

$$\lambda(\mathbb{T}_{W_R}) = \Big(\underbrace{\overbrace{p_0^*}^{d_0=1}, \overbrace{p_1^*,\dots,p_1^*}^{d_1=d},\dots,\overbrace{p_R^*,\dots,p_R^*}^{d_R}}_{\widetilde{R}}, 0, 0, \dots\Big),$$

$$\text{and}\quad \lambda^R(\widehat{T}_n) = \Big(\underbrace{\lambda^R(\widehat{T}_n)_1,\dots,\lambda^R(\widehat{T}_n)_{\widetilde{R}}}_{\widetilde{R}}, 0, 0, \dots\Big),$$

with $\lambda^R(\widehat{T}_n)_1 \ge \cdots \ge \lambda^R(\widehat{T}_n)_{\widetilde{R}}$.
$\rightsquigarrow$ If $p_k^* \neq 0$ for all $0 \le k \le R$, then it is clear that $\sigma^*(\{1,\dots,\widetilde{R}\}) = \{1,\dots,\widetilde{R}\}$. Otherwise, there would exist some $i \in \{1,\dots,\widetilde{R}\}$ such that $\sigma^*(j) \neq i$ for all $j \in \{1,\dots,\widetilde{R}\}$. Hence, we would obtain that $\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) \ge |\lambda(\mathbb{T}_{W_R})_i|^2 \ge (\Delta^G)^2$, which would contradict Eq.(3.21).
$\rightsquigarrow$ If $p_k^* = 0$ for all $0 \le k \le R$, it is clear that we can take $\sigma^* = \mathrm{Id}$.
$\rightsquigarrow$ Otherwise, let us denote $Null$ the list of all indexes $i \in \{1,\dots,\widetilde{R}\}$ such that $\lambda(\mathbb{T}_{W_R})_i = 0$. It holds that $N_0 = |Null| = \sum_{0\le k\le R \text{ s.t. } p_k^*=0} d_k$. We also denote $NoNull$ the complement of $Null$ in $\{1,\dots,\widetilde{R}\}$ (i.e. the list of indexes in $\{1,\dots,\widetilde{R}\}$ that are not in $Null$).
For any $1 \le i \le \widetilde{R}$ such that $\lambda(\mathbb{T}_{W_R})_i \neq 0$, it must exist some $j \in \{1,\dots,\widetilde{R}\}$ such that $\sigma^*(j) = i$. Otherwise, we would have

$$\delta_2^2(\lambda(\mathbb{T}_{W_R}), \lambda^R(\widehat{T}_n)) \ge |\lambda(\mathbb{T}_{W_R})_i|^2 \ge (\Delta^G)^2,$$

which would contradict Eq.(3.21). Hence, we get that

$$(\sigma^*)^{-1}(NoNull) \subset \{1,\dots,\widetilde{R}\}.$$

We deduce that for any $i \in \{1,\dots,\widetilde{R}\}\backslash(\sigma^*)^{-1}(NoNull)$, $\lambda(\mathbb{T}_{W_R})_{\sigma^*(i)} = 0$. Hence, we can define σ^* such that this permutation sends the N_0 indexes in $\{1,\dots,\widetilde{R}\}\backslash(\sigma^*)^{-1}(NoNull)$ to the N_0 indexes in $Null$. Such σ^* still achieves the minimum in Eq.(3.22). In the following, we thus consider that $\sigma^*(\{1,\dots,\widetilde{R}\}) = \{1,\dots,\widetilde{R}\}$.

• Let us recall that the function f^* is defined by

$$\begin{aligned} f^* : \{1, \dots, \widetilde{R}\} &\to \{p_k^*,\ 0 \le k \le R\} \\ i &\mapsto \lambda(\mathbb{T}_{W_R})_{\sigma^*(i)}. \end{aligned}$$

Note that for any $1 \le i \le \widetilde{R}$, $\sigma^*(i) \le \widetilde{R}$ thanks to the previous paragraph. We denote $p^*_{(0)} \ge \cdots \ge p^*_{(R)}$ the ordered sequence of $p_0^*, \dots, p_R^*$ and $d_{(k)}$ is the multiplicity of the eigenvalue $p^*_{(k)}$ of the operator $\mathbb{T}_W$. We show that f^* is such that $f^*(1) = \cdots = f^*(d_{(0)}) = p^*_{(0)}$, $f^*(d_{(0)}+1) = \cdots = f^*(d_{(0)}+d_{(1)}) = p^*_{(1)}$, $f^*(d_{(0)}+d_{(1)}+1) = \cdots = f^*(d_{(0)}+d_{(1)}+d_{(2)}) = p^*_{(2)}, \dots$. This is equivalent to say that the function f^* is non-increasing. If this was not true, it would mean that there exist $1 \le j < i \le \widetilde{R}$ such that $f^*(j) < f^*(i)$. Since $\lambda^R(\widehat{T}_n)_j \ge \lambda^R(\widehat{T}_n)_i$ (because $j < i$), we would get that

$$\begin{aligned} \Delta^G &< f^*(i) - f^*(j) \\ &= \underbrace{f^*(i) - \lambda^R(\widehat{T}_n)_i}_{\le g} + \underbrace{\lambda^R(\widehat{T}_n)_i - \lambda^R(\widehat{T}_n)_j}_{\le 0} + \underbrace{\lambda^R(\widehat{T}_n)_j - f^*(j)}_{\le g} \\ i.e.\quad \Delta^G &\le 2g. \end{aligned}$$

Since we chose g such that $\Delta^G > 4g$, this previous inequality is absurd. This concludes the proof.

3.10.1.2 Proof of Lemma 3.5

We prove our result by induction. In the following, we say that an intermediate state of the HAC algorithm is *valid* if it is still possible to reach state $(\mathcal{S})$ in the next iterations of the algorithm. Stated otherwise, a state is *valid* if it does not exist $1 \le i \ne j \le \widetilde{R}$ such that $f^*(i) \ne f^*(j)$ with $\lambda^R(\widehat{T}_n)_i$ and $\lambda^R(\widehat{T}_n)_j$ in the same cluster. It is obvious that the initial state of the HAC algorithm is *valid* since all eigenvalues are alone in their respective clusters.
Suppose now that we are at iteration $2 \le t \le \widetilde{R} - R - 2$ of the HAC algorithm and that our procedure is *valid* until step t. We are sure that we did not reach a state of type $(\mathcal{S})$ before step t because only the state at depth R from the root of the HAC's tree contains exactly $R+1$ clusters. For any cluster S formed at step t by the HAC algorithm, we denote by abuse of notation $f^*(S) := f^*(i)$ for any i such that $\lambda^R(\widehat{T}_n)_i \in S$ (which is licit since step t is *valid*). By contradiction, assume that the algorithm does not make a valid merging at step $t+1$. This means that the two merged clusters S_a and S_b at step $t+1$ are such that $f^*(S_a) \ne f^*(S_b)$. Since at step t we did not reach a state of type $(\mathcal{S})$, this means that there are two clusters S_i and S_j with $i \ne j$ such that $f^*(S_i) = f^*(S_j)$.
For any $\lambda^R(\widehat{T}_n)_i \in S_i$ and $\lambda^R(\widehat{T}_n)_j \in S_j$,

$$|\lambda^R(\widehat{T}_n)_i - \lambda^R(\widehat{T}_n)_j| \le |\lambda^R(\widehat{T}_n)_i - f^*(S_i)| + \underbrace{|f^*(S_i) - \lambda^R(\widehat{T}_n)_j|}_{=|f^*(S_j) - \lambda^R(\widehat{T}_n)_j|} \le 2g,$$

and for any $\lambda^R(\widehat{T}_n)_a \in S_a$ and $\lambda^R(\widehat{T}_n)_b \in S_b$,

$$\begin{aligned} &|\lambda^R(\widehat{T}_n)_a - \lambda^R(\widehat{T}_n)_b| \\ &\ge -|\lambda^R(\widehat{T}_n)_a - f^*(S_a)| + |f^*(S_a) - \lambda^R(\widehat{T}_n)_b| \\ &\ge |f^*(S_a) - f^*(S_b)| - |\lambda^R(\widehat{T}_n)_a - f^*(S_a)| - |\lambda^R(\widehat{T}_n)_b - f^*(S_b)| \\ &\ge \Delta^G - 2g. \end{aligned}$$

Since we chose $\Delta^G > 4g$, we get

$$d_c(S_a, S_b) > d_c(S_i, S_j).$$

This is a contradiction since at step t, the HAC algorithm merges the two clusters with the smallest complete linkage distance. Hence, the algorithm performs a valid merging at step $t+1$.
We proved that a state of type $(\mathcal{S})$ is reached by the HAC algorithm with complete linkage at iteration $\widetilde{R} - R - 1$. Since $d \ge 3$, it holds $d_0 < d_1 < d_2 < \dots$ and since the SCCHEi starts by selecting the cluster of size d_0 in the tree as close as possible to the root, we get $\mathcal{C}_{d_0} = \widehat{\mathcal{C}}_{d_0}$. Continuing the process of

the "for loop" in the SCCHEi algorithm, the SCCHEi algorithm then selects the cluster of size d_1 in the remaining tree (where we removed all eigenvalues in $\widehat{\mathcal{C}}_{d_0}$ in the tree of the HAC). Hence, the SCCHEi algorithm sets $\mathcal{C}_{d_1} = \widehat{\mathcal{C}}_{d_1}$. Following this procedure, it is straightforward to see that the SCCHEi returns the partition $\mathcal{C}_{d_0} = \widehat{\mathcal{C}}_{d_0}, \ldots, \mathcal{C}_{d_R} = \widehat{\mathcal{C}}_{d_R}$.

3.10.2 Concentration inequality for U-statistics with Markov chains

In this section, we present briefly the main result from Chapter 4 of this thesis: a concentration inequality for a U-statistic of the Markov chain $(X_i)_{i\geq 1}$. This concentration inequality is a key result to prove Theorem 3.2. In the first subsection, we remind the assumptions made on the Markovian dynamic, namely Assumption A.

3.10.2.1 Assumptions and notations for the Markov chain

Let us recall that Assumption A states that the latitude function $f_{\mathcal{L}}$ is such that $\|f_{\mathcal{L}}\|_\infty < \infty$ and makes the chain $(X_i)_{i\geq 1}$ uniformly ergodic. Assumption A guarantees in particular that there exists $\delta_M > 0$ such that

$$\forall x \in \mathbb{S}^{d-1}, \forall A \in \mathcal{B}(\mathbb{S}^{d-1}), \quad P(x, A) \leq \delta_M \nu(A),$$

for some probability measure ν (e.g. the uniform measure on the sphere π).
In Section 3.9.2, we provide a sufficient condition on the latitude function $f_{\mathcal{L}}$ ensuring the uniform ergodicity of the chain with associated constants $L > 0$ and $0 < \rho < 1$ (cf. Definition A.8). In Section 3.9.3, we explain why Assumption A ensures that the Markov chain $(X_i)_{i\geq 1}$ has an absolute spectral gap (cf. Definition A.10) and we show that this absolute spectral gap is equal to 1.

3.10.2.2 Concentration inequality of U-statistic for Markov chain

One key result to prove Theorem 3.2 is the concentration of the following U-statistic

$$U_{\text{stat}}(n) = \frac{1}{n^2} \sum_{1\leq i<j\leq n} \left[(W - W_R)^2(X_i, X_j) - \|W - W_R\|_2^2\right].$$

Note that $\|W - W_R\|_2^2$ corresponds to the expectation of the kernel $(W - W_R)^2(\cdot, \cdot)$ under the uniform distribution on $\mathbb{S}^{d-1}$ which is known to be the unique stationary distribution π of the Markov chain $(X_i)_{i\geq 1}$ (cf. Section 3.9). More precisely, for any $x \in \mathbb{S}^{d-1}$, it holds

$$\|W - W_R\|_2^2 = \mathbb{E}_{X\sim\pi}[(W - W_R)^2(x, X)] = \mathbb{E}_{(X,X')\sim\pi\otimes\pi}[(W - W_R)^2(X, X')],$$

see Lemma 3.1 for a proof. Applying Theorem 4.3 from Chapter 4 in a our framework leads to the following result.

Lemma 3.19. *Let us consider $\gamma \in (0, 1)$ satisfying $\log(e\log(n)/\gamma) \leq n$. Then it holds with probability at least $1 - \gamma$,*

$$U_{\text{stat}}(n) \leq M \frac{\|\mathbf{p} - \mathbf{p}_R\|_\infty^2 \log n}{n} \log(e\log(n)/\gamma),$$

where $M > 0$ only depends on constants related to the Markov chain $(X_i)_{i\geq 1}$.

3.10.3 Proof of Theorem 3.2

The proof of Theorem 3.2 mainly lies in the following result which is proved in Section 3.10.3.1. Coupling the convergence of the spectrum of the matrix of probability T_n with a concentration result on the spectral norm of random matrices with independent entries (cf. Bandeira and van Handel [2016]), we show the convergence in metric δ_2 of the spectrum of $\widehat{T}_n$ towards the spectrum of the integral operator $\mathbb{T}_W$.

Theorem 3.20. *Let us consider $\gamma \in (0, 1)$ satisfying $\log(e\log(n)/\gamma) \leq n/(13\widetilde{R})$. Then it holds with probability*

at least $1-\gamma$,

$$\begin{aligned} & \delta_2\left(\lambda(\mathbb{T}_W), \lambda(T_n)\right) \\ \leq \; & 2\|\mathbf{p}-\mathbf{p}_R\|_2 + 8\sqrt{\frac{\widetilde{R}}{n}\ln(e/\gamma)} + M\|\mathbf{p}-\mathbf{p}_R\|_\infty \sqrt{\frac{\log n}{n}}\left(\log(e\log(n)/\gamma)\right)^{1/2}, \end{aligned}$$

where $M>0$ *only depends on constants related to the Markov chain* $(X_i)_{i\geq 1}$ *(cf. Lemma 3.19).*

First part of the proof for Theorem 3.2 We start by establishing the convergence rate for $\delta_2\left(\lambda(\mathbb{T}_W), \lambda(T_n)\right)$. We keep notations of Theorem 3.20. Let us consider $\gamma \in (0,1)$ satisfying $\log(e\log(n)/\gamma) \leq (n/(13\widetilde{R}))$, and assume that $p \in Z^s_{w_\beta}((-1,1))$ with $s>0$.
Let us define the event

$$\begin{aligned} \Omega(\gamma) := \Bigg\{ & \delta_2\left(\lambda(\mathbb{T}_W), \lambda(T_n)\right) \leq 2\|\mathbf{p}-\mathbf{p}_R\|_2 + 8\sqrt{\frac{\widetilde{R}}{n}\ln(e/\gamma)} \\ & + M\|\mathbf{p}-\mathbf{p}_R\|_\infty \sqrt{\frac{\log n}{n}}\left(\log(e\log(n)/\gamma)\right)^{1/2} \Bigg\}. \end{aligned}$$

Using Theorem 3.20, it holds $\mathbb{P}\left(\Omega(\gamma)\right) \geq 1-\gamma$. Remarking further that

$$\delta_2\left(\lambda(\mathbb{T}_W), \lambda(T_n)\right) \leq \delta_2\left(\lambda(\mathbb{T}_W), 0\right) + \delta_2\left(0, \lambda(T_n)\right) \leq \|\mathbf{p}\|_2 + \sqrt{n} \leq \sqrt{2} + \sqrt{n},$$

we have

$$\begin{aligned} & \mathbb{E}[\delta_2^2(\lambda(\mathbb{T}_W), \lambda(T_n))] \\ = \; & \mathbb{E}[\delta_2^2(\lambda(\mathbb{T}_W), \lambda(T_n))\mathbb{1}_{\Omega(\gamma)}] + (1+\sqrt{2})^2 n\mathbb{P}(\Omega(\gamma)^c) \\ \leq \; & c\|\mathbf{p}-\mathbf{p}_R\|_2^2 + c\frac{\widetilde{R}}{n}\log(e/\gamma) + c\|\mathbf{p}-\mathbf{p}_R\|_\infty^2 \frac{\log n}{n}\log(e\log(n)/\gamma) \\ & + (1+\sqrt{2})^2 n\gamma, \end{aligned}$$

where $c>0$ is a constant that does not depend on R, d nor n. Since for some constant $C(\mathbf{p}, s, d) > 0$ (depending only on $\mathbf{p}$, s and d)

$$\|\mathbf{p}-\mathbf{p}_R\|_2^2 = \sum_{k>R} (p_k^*)^2 d_k \frac{(1+k(k+2\beta))^s}{(1+k(k+2\beta))^s} \leq C(\mathbf{p}, s, d) R^{-2s}, \tag{3.23}$$

and since

$$\widetilde{R} = O(R^{d-1}), \tag{3.24}$$

we have choosing $\gamma = 1/n^2$

$$\mathbb{E}[\delta_2^2(\lambda(\mathbb{T}_W), \lambda(T_n))] \leq D'\left[R^{-2s} + R^{d-1}\frac{\log(n)}{n} + \|\mathbf{p}-\mathbf{p}_R\|_\infty^2 \frac{\log^2(n)}{n}\right], \tag{3.25}$$

where $D'>0$ is a constant independent of n and R. Let us show that choosing $R = \lfloor \left(n/\log^2(n)\right)^{\frac{1}{2s+d-1}} \rfloor$ concludes the proof. Since $\|G_k^\beta\|_\infty = G_k^\beta(1) = d_k/c_k$, we get that

$$\|\mathbf{p}_R\|_\infty \leq \sum_{k=0}^{R} |p_k^*| c_k G_k^\beta(1) = \sum_{k=0}^{R} |p_k^*| d_k \leq \sqrt{\widetilde{R}}\|\mathbf{p}_R\|_2,$$

and using Eq.(3.30), we deduce that

$$\|\mathbf{p}-\mathbf{p}_R\|_\infty \leq \|\mathbf{p}\|_\infty + \|\mathbf{p}_R\|_\infty \leq 1 + \sqrt{2\widetilde{R}}. \tag{3.26}$$

Hence, Eq.(3.25) becomes

$$\mathbb{E}[\delta_2^2(\lambda(\mathbb{T}_W),\lambda(T_n))] \leq D''\left[R^{-2s} + R^{d-1}\frac{\log(n)}{n} + \widetilde{R}\frac{\log^2(n)}{n}\right],$$

where D'' is a constant that does not depend on n nor R. Choosing $R = \lfloor \left(n/\log^2(n)\right)^{\frac{1}{2s+d-1}} \rfloor$ and using Eq.(3.24) we get

$$\begin{aligned}
& \mathbb{E}[\delta_2^2(\lambda(\mathbb{T}_W),\lambda(T_n))] \\
\leq\ & D''\left[\left(\frac{n}{\log^2(n)}\right)^{\frac{-2s}{2s+d-1}} + 2\left(\frac{n}{\log^2(n)}\right)^{\frac{d-1}{2s+d-1}}\frac{\log^2(n)}{n}\right] \\
\leq\ & 3D''\left(\frac{n}{\log^2(n)}\right)^{\frac{-2s}{2s+d-1}}.
\end{aligned}$$

Second part of the proof for Theorem 3.2 Let us recall that in the statement of Theorem 3.2, $\lambda^{R_{opt}}(\widehat{T}_n)$ is the sequence of the $\widetilde{R}_{opt}$ first eigenvalues (sorted in decreasing absolute values) of the matrix $\widehat{T}_n$ where R_{opt} is the value of the parameter R leading to the optimal bias-variance trade off, namely

$$\lambda^{R_{opt}}(\widehat{T}_n) = (\hat{\lambda}_1,\dots,\hat{\lambda}_{\widetilde{R}_{opt}},0,0,\dots).$$

From the computations of the first part of the proof, we know that $R_{opt} = \lfloor \left(n/\log^2(n)\right)^{\frac{1}{2s+d-1}} \rfloor$. That corresponds to the situation where we choose optimally R and it is in practice possible to approximate this best model dimension using e.g. the slope heuristic. Therefore, $\delta_2\left(\lambda(\mathbb{T}_W),\lambda^{R_{opt}}(\widehat{T}_n)\right)$ is the quantity of interest since it represents the distance between the eigenvalues used to built our estimates $(\hat{p}_k)_k$ and the true spectrum of the envelope function $\mathbf{p}$. Since $\widetilde{R} = \mathcal{O}\left(R^{d-1}\right)$ for all integer $R \geq 0$, we have $\widetilde{R}_{opt} = \mathcal{O}\left((n/\log^2(n))^{\frac{d-1}{2s+d-1}}\right)$. We deduce that for n large enough $2\widetilde{R}_{opt} \leq n$ and using [De Castro et al., 2019, Proposition 15] we obtain

$$\begin{aligned}
& \delta_2\left(\lambda^{R_{opt}}(\widehat{T}_n),\lambda(\mathbb{T}_{W_{R_{opt}}})\right) \\
\leq\ & \delta_2\left(\lambda(T_n),\lambda(\mathbb{T}_{W_{R_{opt}}})\right) + \sqrt{2\widetilde{R}_{opt}}\|\widehat{T}_n - T_n\| \\
\leq\ & \delta_2\left(\lambda(T_n),\lambda(\mathbb{T}_W)\right) + \delta_2\left(\lambda(\mathbb{T}_W),\lambda(\mathbb{T}_{W_{R_{opt}}})\right) + \sqrt{2\widetilde{R}_{opt}}\|\widehat{T}_n - T_n\|, \qquad (3.27)
\end{aligned}$$

where $\lambda(\mathbb{T}_{W_{R_{opt}}}) = (\lambda_1^*,\dots,\lambda^*_{\widetilde{R}_{opt}},0,0,\dots)$. Let us consider $\gamma \in (0,1)$. Using Theorem 3.20, we know that with probability at least $1-\gamma$ it holds for n large enough

$$\begin{aligned}
\delta_2\left(\lambda(T_n),\lambda(\mathbb{T}_W)\right) \leq\ & 2\|\mathbf{p} - \mathbf{p}_{R_{opt}}\|_2 + 8\sqrt{\frac{\widetilde{R}_{opt}}{n}\ln(e/\gamma)} \\
& + M\|\mathbf{p} - \mathbf{p}_{R_{opt}}\|_\infty\sqrt{\frac{\log n}{n}}\left(\log(e\log(n)/\gamma)\right)^{1/2}.
\end{aligned}$$

Using Eq.(3.23), Eq.(3.26) and the fact that $\widetilde{R} = \mathcal{O}(R^{d-1})$, it holds with probability at least $1 - 1/n^2$,

$$\begin{aligned}
\delta_2^2\left(\lambda(T_n),\lambda(\mathbb{T}_W)\right) &\leq c\left[R_{opt}^{-2s} + R_{opt}^{d-1}\frac{\log n}{n} + MR_{opt}^{d-1}\frac{\log^2 n}{n}\right] \\
&\leq (M')^2(n/\log^2 n)^{\frac{-2s}{2s+d-1}},
\end{aligned}$$

where $c > 0$ is a numerical constant and $M' > 0$ depends on constants related to the Markov chain $(X_i)_{i\geq 1}$ (see Theorem 3.20 for details). Moreover,

$$\begin{aligned}\delta_2^2\left(\lambda(\mathbb{T}_W), \lambda(\mathbb{T}_{W_{R_{opt}}})\right) &= \|\mathbf{p} - \mathbf{p}_{R_{opt}}\|_2^2 \\ &\leq C(\mathbf{p}, s, d) R_{opt}^{-2s} = \mathcal{O}\left((n/\log^2 n)^{\frac{-2s}{2s+d-1}}\right), \end{aligned} \tag{3.28}$$

where we used Eq.(3.23). Finally, using the concentration of spectral norm for random matrices with independent entries from Bandeira and van Handel [2016], there exists a universal constant $C_0 > 0$ such that conditionally on $(X_i)_{i\geq 1}$, it holds with probability at least $1 - 1/n^2$,

$$\|T_n - \widehat{T}_n\| \leq \frac{3}{\sqrt{2n}} + C_0 \frac{\sqrt{\log(n^3)}}{n}.$$

Using again $\widetilde{R} = \mathcal{O}(R^{d-1})$, this implies that for n large enough, it holds conditionally on $(X_i)_{i\geq 1}$ with probability at least $1 - 1/n^2$,

$$\sqrt{2\widetilde{R}_{opt}}\|T_n - \widehat{T}_n\| \leq D(n/\log^2 n)^{\frac{-s}{2s+d-1}},$$

where $D > 0$ is a numerical constant. From Eq.(3.27), we deduce that $\mathbb{P}(\Omega) \geq 1 - 2/n^2$ where the event Ω is defined by

$$\Omega = \left\{\delta_2^2\left(\lambda^{R_{opt}}(\widehat{T}_n), \lambda(\mathbb{T}_{W_{R_{opt}}})\right) \leq \left(C(\mathbf{p}, s, d)^{1/2} + D + M'\right)^2 (n/\log^2 n)^{\frac{-2s}{2s+d-1}}\right\}.$$

Remarking finally that

$$\begin{aligned}\delta_2\left(\lambda^{R_{opt}}(\widehat{T}_n), \lambda(\mathbb{T}_{W_{R_{opt}}})\right) &\leq \delta_2\left(\lambda(\mathbb{T}_{W_{R_{opt}}}), 0\right) + \delta_2\left(0, \lambda(\widehat{T}_n)\right) \\ &\leq \|\mathbf{p}\|_2 + \sqrt{n} \leq \sqrt{2} + \sqrt{n},\end{aligned}$$

we obtain

$$\begin{aligned}& \mathbb{E}\left[\delta_2^2\left(\lambda^{R_{opt}}(\widehat{T}_n), \lambda(\mathbb{T}_{W_{R_{opt}}})\right)\right] \\ \leq\ & \mathbb{E}\left[\delta_2^2\left(\lambda^{R_{opt}}(\widehat{T}_n), \lambda(\mathbb{T}_{W_{R_{opt}}})\right) \mid \Omega\right] + \mathbb{P}(\Omega^c)(\sqrt{2} + \sqrt{n})^2 \\ \leq\ & \left(C(\mathbf{p}, s, d)^{1/2} + D + M'\right)^2 (n/\log^2 n)^{\frac{-2s}{2s+d-1}} + 2\frac{(\sqrt{2} + \sqrt{n})^2}{n^2} \\ =\ & \mathcal{O}\left((n/\log^2 n)^{\frac{-2s}{2s+d-1}}\right). \end{aligned} \tag{3.29}$$

Using the triangle inequality, Eq.(3.28) and Eq.(3.29) lead to

$$\begin{aligned}&\mathbb{E}\left[\delta_2^2\left(\lambda^{R_{opt}}(\widehat{T}_n), \lambda(\mathbb{T}_W)\right)\right] \\ &\quad \leq 3\mathbb{E}\left[\delta_2^2\left(\lambda^{R_{opt}}(\widehat{T}_n), \lambda(T_{W_{R_{opt}}})\right)\right] + 3\delta_2^2\left(\lambda(T_{W_{R_{opt}}}), \lambda(T_W)\right) \\ &\quad = \mathcal{O}\left((n/\log^2 n)^{\frac{-2s}{2s+d-1}}\right),\end{aligned}$$

which concludes the proof of Theorem 3.2.

3.10.3.1 Proof of Theorem 3.20

We follow the same sketch of proof as in De Castro et al. [2019]. Let $R \geq 1$ and define,

$$\begin{aligned}\Phi_{k,l} &= \frac{1}{\sqrt{n}}\left[Y_{k,l}(X_1), \ldots, Y_{k,l}(X_n)\right] \in \mathbb{R}^n, \\ E_{R,n} &= \left(\langle \Phi_{k,l}, \Phi_{k',l'}\rangle - \delta_{(k,l),(k',l')}\right)_{(k,k')\in[R],\ l\in\{1,\ldots,d_k\},\ l'\in\{1\ldots,d_{k'}\}} \in \mathbb{R}^{\widetilde{R}\times\widetilde{R}},\end{aligned}$$

$$
\begin{aligned}
X_{R,n} &= [\Phi_{0,1}, \Phi_{1,1}, \Phi_{1,2}, \dots, \Phi_{R,d_R}] \in \mathbb{R}^{n\times\widetilde{R}},\\
A_{R,n} &= \left(X_{R,n}^\top X_{R,n}\right)^{1/2} \text{ with } A_{R,n}^2 = \mathrm{Id}_{\widetilde{R}} + E_{R,n},\\
K_R &= \mathrm{Diag}(\lambda_1(\mathbb{T}_W), \dots, \lambda_{\widetilde{R}}(\mathbb{T}_W)),\\
T_{R,n} &= \sum_{k=0}^{R} p_k^* \sum_{l=1}^{d_k} \Phi_{k,l}(\Phi_{k,l})^\top = X_{R,n} K_R X_{R,n}^\top \in \mathbb{R}^{n\times n}\\
\tilde{T}_{R,n} &= ((1-\delta_{i,j})T_{R,n})_{i,j\in[n]} \in \mathbb{R}^{n\times n},\\
T_{R,n}^* &= A_{R,n} K_R A_{R,n}^\top \in \mathbb{R}^{\widetilde{R}\times\widetilde{R}},\\
W_R(x,y) &= \sum_{k=0}^{R} p_k^* \sum_{l=1}^{d_k} Y_{k,l}(x) Y_{k,l}(y).
\end{aligned}
$$

It holds

$$
\delta_2(\lambda(\mathbb{T}_W), \lambda(\mathbb{T}_{W_R})) = \left(\sum_{k>R} d_k (p_k^*)^2\right)^{1/2}.
$$

We point out the equality between spectra of the operator $\mathbb{T}_{W_R}$ and the matrix K_R. Using the SVD decomposition of $X_{R,n}$, one can also easily prove that $\lambda(T_{R,n}) = \lambda(T_{R,n}^*)$. We deduce that

$$
\begin{aligned}
\delta_2\left(\lambda(\mathbb{T}_{W_R}), \lambda(T_{R,n})\right) &= \delta_2\left(\lambda(K_R), \lambda(T_{R,n}^*)\right)\\
&\le \|T_{R,n}^* - K_R\|_F\\
&= \|A_{R,n} K_R A_{R,n} - K_R\|_F,
\end{aligned}
$$

with the Hoffman-Wielandt inequality. Using equation (4.8) at (Koltchinskii and Giné [2000] p.127) gives

$$
\delta_2\left(\lambda(\mathbb{T}_{W_R}), \lambda(T_{R,n})\right) \le \sqrt{2}\|K_R\|_F \|E_{R,n}\| = \sqrt{2}\|W_R\|_2 \|E_{R,n}\|.
$$

Using again the Hoffman-Wielandt inequality we get

$$
\delta_2(\lambda(T_{R,n}), \lambda(\tilde{T}_{R,n})) \le \|\tilde{T}_{R,n} - T_{R,n}\|_F = \left[\frac{1}{n^2}\sum_{i=1}^{n} W_R(X_i, X_i)^2\right]^{1/2},
$$

and

$$
\delta_2\left(\lambda(\tilde{T}_{R,n}), \lambda(T_n)\right) \le \|\tilde{T}_{R,n} - T_n\|_F = \left[\frac{1}{n^2}\sum_{i\neq j} (W - W_R)^2(X_i, X_j)\right]^{1/2}.
$$

Now, we invoke Lemmas 3.19, 3.21 and 3.22 to conclude the proof. The proofs of these last two lemmas are provided in Section 3.10.3.2 and Section 3.10.3.3 respectively.

Lemma 3.21. *Let us consider $\gamma > 0$ and assume that $13\widetilde{R}\ln(e/\gamma) \le n$. Then it holds with probability at least $1-\gamma$*

$$
\|E_{R,n}\| \le 4\sqrt{\frac{\widetilde{R}}{n}\ln(2/\gamma)}.
$$

Lemma 3.22. *Let $R \ge 1$. We have*

$$
\frac{1}{n^2}\sum_{i=1}^{n} W_R(X_i, X_i)^2 = \frac{1}{n}\left(\sum_{k=0}^{R} p_k^* d_k\right)^2.
$$

For any $\gamma \in (0,1)$ with $\log(e\log(n)/\gamma) \leq (n/(13\widetilde{R}))$, it holds with probability at least $1-\gamma$,

$$\begin{aligned}
& \delta_2\left(\lambda(\mathbb{T}_W), \lambda(T_n)\right) \\
\leq\ & \delta_2\left(\lambda(\mathbb{T}_W), \lambda(\mathbb{T}_{W_R})\right) + \delta_2\left(\lambda(\mathbb{T}_{W_R}), \lambda(T_{R,n})\right) + \delta_2\left(\lambda(T_{R,n}), \lambda(\tilde{T}_{R,n})\right) \\
& + \delta_2\left(\lambda(\tilde{T}_{R,n}), \lambda(T_n)\right) \\
\leq\ & 4\sqrt{\frac{\widetilde{R}}{n}\ln(2/\gamma)} + \sqrt{2}\left(\sum_{k=0}^{R} d_k (p_k^*)^2\right)^{1/2} + \frac{1}{\sqrt{n}}\left|\sum_{k=0}^{R} p_k^* d_k\right| + 2\|\mathbf{p}-\mathbf{p}_R\|_2 \\
& + M\|\mathbf{p}-\mathbf{p}_R\|_\infty \sqrt{\frac{\log n}{n}}\left(\log(e\log(n)/\gamma)\right)^{1/2},
\end{aligned}$$

where $M > 0$ depends only on constants related to the Markov chain $(X_i)_{i\geq 1}$. Now remark that

$$\left|\sum_{k=0}^{R} p_k^* d_k\right| \leq \left(\sum_{k=0}^{R} d_k\right)^{1/2}\left(\sum_{k=0}^{R} d_k (p_k^*)^2\right)^{1/2} = \sqrt{\widetilde{R}}\|\mathbf{p}_R\|_2,$$

and that

$$\|\mathbf{p}_R\|_2^2 \leq \|\mathbf{p}\|_2^2 \leq 2, \tag{3.30}$$

because $\mathbf{p}_R$ is the orthogonal projection of $\mathbf{p}$, and $|\mathbf{p}| \leq 1$. We deduce that

$$\begin{aligned}
& \delta_2\left(\lambda(\mathbb{T}_W), \lambda(T_n)\right) \\
\leq\ & 2\|\mathbf{p}-\mathbf{p}_R\|_2 + 4\sqrt{\frac{\widetilde{R}}{n}\ln(2/\gamma)} + \sqrt{\frac{2\widetilde{R}}{n}} \\
& + M\|\mathbf{p}-\mathbf{p}_R\|_\infty \sqrt{\frac{\log n}{n}}\left(\log(e\log(n)/\gamma)\right)^{1/2} \\
\leq\ & 2\|\mathbf{p}-\mathbf{p}_R\|_2 + 8\sqrt{\frac{\widetilde{R}}{n}\ln(e/\gamma)} \\
& + M\|\mathbf{p}-\mathbf{p}_R\|_\infty \sqrt{\frac{\log n}{n}}\left(\log(e\log(n)/\gamma)\right)^{1/2}.
\end{aligned}$$

3.10.3.2 Proof of Lemma 3.21

Observe that $nE_{R,n} = \sum_{i=1}^{n}\left(Z_i Z_i^\top - \mathrm{Id}_{\widetilde{R}}\right)$ where for all $i \in [n]$, $Z_i \in \mathbb{R}^{\widetilde{R}}$ is defined by

$$\begin{aligned}
Z_i := Z(X_i) := \big(& Y_{0,1}(X_i), Y_{1,1}(X_i), Y_{1,2}(X_i), \ldots, Y_{1,d_1}(X_i), \ldots, \\
& Y_{R,1}(X_i), \ldots, Y_{R,d_R}(X_i)\big).
\end{aligned}$$

By definition of the spectral norm for a Hermitian matrix,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top - \mathrm{Id}_{\widetilde{R}}\right\| = \max_{x,\ \|x\|_2=1}\left|x^\top\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top\right)x - 1\right|.$$

We use a covering set argument based on the following Lemma.

Lemma 3.23. *(cf. [Gilles, 1989, Lemma 4.10])*
Let us consider an integer $D \geq 2$. For any $\epsilon_0 > 0$, there exists a set $Q \subset \mathbb{S}^{D-1}$ of cardinality at most $(1+2/\epsilon_0)^D$ such that

$$\forall \alpha \in \mathbb{S}^{D-1}, \quad \exists q \in Q, \quad \|\alpha - q\|_2 \leq \epsilon_0.$$

We consider Q the set given by Lemma 3.23 with $D = d$ and $\epsilon_0 \in (0, 1/2)$. Let us define $x_0 \in \mathbb{S}^{d-1}$ such

that $|x_0^\top E_{R,n} x_0| = \|E_{R,n}\|$ and $q_0 \in Q$ such that $\|x_0 - q_0\|_2 \leq \epsilon_0$. Then,

$$\begin{aligned}
|x_0^\top E_{R,n} x_0| - |q_0^\top E_{R,n} q_0| &\leq |x_0^\top E_{R,n} x_0 - q_0^\top E_{R,n} q_0| \text{ (by triangle inequality)} \\
&= |x_0^\top E_{R,n}(x_0 - q_0) - (q_0 - x_0)^\top E_{R,n} q_0| \\
&\leq \|x_0\|_2 \|E_{R,n}\| \|x_0 - q_0\|_2 + \|q_0 - x_0\|_2 \|E_{R,n}\| \|q_0\|_2 \\
&\leq 2\epsilon_0 \|E_{R,n}\|.
\end{aligned}$$

which leads to

$$|x_0^\top E_{R,n} x_0| = \|E_{R,n}\| \leq |q_0^\top E_{R,n} q_0| + 2\epsilon_0 \|E_{R,n}\|.$$

Hence,

$$\|E_{R,n}\| \leq \frac{1}{1 - 2\epsilon_0} \max_{q \in Q} |q^\top E_{R,n} q|.$$

We introduce for any $q \in Q$ the function

$$F_q : x = (x_1, \ldots, x_n) \mapsto \frac{1}{n} \sum_{i=1}^n q^\top \left(Z_i Z_i^\top - 1 \right) q := \frac{1}{n} \sum_{i=1}^n f_q(x_i),$$

where $f_q(x) = q^\top \left(Z(x) Z(x)^\top - 1 \right) q$.
Let us consider $t > 0$. We want to apply Bernstein's inequality for Markov chains from [Jiang et al., 2018, Theorem 1.1]. In the following, we denote $\mathbb{E}_\pi[\cdot]$ the expectation with respect to the measure π. We remark that $\mathbb{E}_\pi[f_q(X)] = 0$ and that $\|f_q\|_\infty \leq \widetilde{R} - 1$. For all $m \in [\widetilde{R}]$, we denote $\phi_m = Y_{r,l}$ with $r \in \{0, \ldots, R\}$ and $l \in [d_r]$ such that $m = l + \sum_{i=0}^r d_i - 1$. Then, for any $x \in \mathbb{S}^{d-1}$, and for all $k, l \in [\widetilde{R}]$, $\left((Z(x)^\top Z(x))^2 \right)_{k,l} = \sum_{m=1}^{\widetilde{R}} \phi_l(x) \phi_m(x)^2 \phi_k(x) = \widetilde{R} \phi_l(x) \phi_k(x) = \widetilde{R} \left(Z(x) Z(x)^\top \right)_{k,l}$ where we used [Dai and Xu, 2013, Eq.(1.2.9)]. We deduce that

$$\begin{aligned}
\mathbb{E}_\pi[f_q(X)^2] &= \mathbb{E}_\pi[q^\top Z(X) Z(X)^\top q q^\top Z(x) Z(x)^\top q] - 2\mathbb{E}_\pi[q^\top Z(X) Z(X)^\top q] + 1 \\
&= \mathbb{E}_\pi[q^\top \underbrace{(Z(X) Z(X)^\top)^2}_{= \widetilde{R} \cdot Z(X) Z(X)^\top} q] - 2 q^\top \underbrace{\mathbb{E}_\pi[Z(X) Z(X)^\top]}_{= \mathrm{Id}} q + 1 \\
&= \widetilde{R} \cdot q^\top \mathbb{E}_\pi[Z(X) Z(X)^\top] q - 1 \\
&= \widetilde{R} - 1.
\end{aligned}$$

Using that the Markov chain $(X_i)_{i \geq 1}$ has an absolute spectral gap equals to 1 (cf. Section 3.9.3), we get from [Jiang et al., 2018, Eq. (1.6)] that

$$\mathbb{P}\left(|F_q(X)| \geq t \right) = \mathbb{P}\left(|q^\top E_{R,n} q| \geq t \right) \leq 2 \exp\left(\frac{-nt^2}{4(\widetilde{R} - 1) + 10(\widetilde{R} - 1)t} \right),$$

which leads to

$$\begin{aligned}
\mathbb{P}\left(\max_{q \in Q} |q^\top E_{R,n} q| \geq t \right) &\leq \mathbb{P}\left(\bigcup_{q \in Q} |q^\top E_{R,n} q| \geq t \right) \\
&\leq 2 \exp\left(\frac{-nt^2/(\widetilde{R} - 1)}{4 + 10t} \right) (1 + 2/\epsilon_0)^{\widetilde{R}}.
\end{aligned}$$

Choosing $\epsilon_0 = 2 \left(\exp\left(\frac{nt^2/2}{(\widetilde{R}-1)\widetilde{R}(4+10t)} \right) - 1 \right)^{-1}$ in order to satisfy $(1 + 2/\epsilon_0)^{\widetilde{R}} = \exp(nt^2(\widetilde{R} - 1)^{-1}(4 + 10t)^{-1}/2)$, we get

$$\mathbb{P}\left(\max_{q \in Q} |q^\top E_{R,n} q| \geq t \right) \leq 2 \exp\left(\frac{-nt^2}{(\widetilde{R} - 1)(8 + 20t)} \right).$$

We deduce that if $\frac{25}{2}\ln(2/\alpha)\widetilde{R} \leq n$, it holds with probability at least $1-\alpha$,

$$\max_{q\in Q} |q^\top E_{R,n} q| \leq 16\sqrt{\frac{\widetilde{R}}{n}\ln(2/\alpha)}.$$

Assuming that $200\ln(7)\widetilde{R}^3\ln(2/\alpha) \leq n^3$ in order to have $1/(1-2\epsilon_0) \leq 4$, it holds with probability at least $1-\alpha$

$$\|E_{R,n}\| \leq \frac{1}{1-2\epsilon_0}\max_{q\in Q}|q^\top E_{R,n}q| \leq 4\sqrt{\frac{\widetilde{R}}{n}\ln(2/\alpha)}.$$

3.10.3.3 Proof of Lemma 3.22

Reminding that for all $x \in \mathbb{S}^{d-1}$ and for all $k \geq 0$, $\sum_{l=1}^{d_k} Y_{k,l}(x)^2 = d_k$ (cf. Corollary 1.2.7 from Dai and Xu [2013]), we get

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n W_R(X_i,X_i)^2 &= \frac{1}{n^2}\sum_{i=1}^n\left(\sum_{k=0}^R p_k^*\sum_{l=1}^{d_k} Y_{k,l}(X_i)^2\right)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n\left(\sum_{k=0}^R p_k^* d_k\right)^2 \\
&= \frac{1}{n}\left(\sum_{k=0}^R p_k^* d_k\right)^2.
\end{aligned}$$

3.10.4 Proof of Theorem 3.8

Proposition 3.24 is the counterpart of Proposition 1 in Araya and De Castro [2019] in our dependent framework. This result is the cornerstone of Theorem 3.8 and is proved in Section 3.10.4.1.

Proposition 3.24. *We assume that $\Delta^* > 0$. Let us consider $\gamma > 0$ and define the event*

$$\mathcal{E} := \left\{\delta_2(\lambda(T_n),\lambda(\mathbb{T}_W)) \vee \frac{2^{\frac{9}{2}}\sqrt{d}}{\Delta^*}\|T_n - \widehat{T}_n\| \leq \frac{\Delta^*}{4}\right\}.$$

Then for n large enough,

$$\mathbb{P}(\mathcal{E}) \geq 1-\gamma/2.$$

Moreover, on the event $\mathcal{E}$, there exists one and only one set Λ_1, consisting of d eigenvalues of $\widehat{T}_n$, whose diameter is smaller that $\Delta^*/2$ and whose distance to the rest of the spectrum of $\widehat{T}_n$ is at least $\Delta^*/2$. Furthermore, on the event $\mathcal{E}$, the algorithm HEiC returns the matrix $\hat{G} = \frac{1}{d}\hat{V}\hat{V}^\top$, where $\hat{V}$ has by columns the eigenvectors corresponding to the eigenvalues in Λ_1.

In the following, we work on the event $\mathcal{E}$. Let us consider $\gamma \in (0,1)$.
We choose $R = (n/\log^2 n)^{\frac{1}{2s+d-1}}$. Reminding that W_R is the rank R approximation of W, the Gram matrix associated with the kernel W_R is

$$T_{R,n} = \sum_{k=0}^R p_k^*\sum_{l=1}^{d_k}\Phi_{k,l}(\Phi_{k,l})^\top = X_{R,n}K_R X_{R,n}^\top \in \mathbb{R}^{n\times n}$$

where

$$\begin{aligned}
\Phi_{k,l} &= \frac{1}{\sqrt{n}}[Y_{k,l}(X_1),\dots,Y_{k,l}(X_n)] \in \mathbb{R}^n, \\
X_{R,n} &= [\Phi_{0,1},\Phi_{1,1},\Phi_{1,2},\dots,\Phi_{R,d_R}] \in \mathbb{R}^{n\times\widetilde{R}} \text{ and} \\
K_R &= \mathrm{Diag}(\lambda_1(\mathbb{T}_W),\dots,\lambda_{\widetilde{R}}(\mathbb{T}_W)).
\end{aligned}$$

Let us denote now $\tilde{V}$ (resp. $\tilde{V}_R$) the orthonormal matrix formed by the eigenvectors of the matrix T_n (resp. $T_{R,n}$). We have the following eigenvalue decompositions

$$T_n = \tilde{V}\Lambda\tilde{V}^\top \text{ and } T_{R,n} = \tilde{V}_R\Lambda_R\tilde{V}_R^\top,$$

where $\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$ are the eigenvalues of the matrix T_n and where $\Lambda_R = (p_0^*, p_1^*, \dots, p_1^*, \dots, p_R^*, \dots, p_R^*, 0, \dots, 0) \in \mathbb{R}^n$ where each p_k^* has multiplicity d_k. Then, we note by $V \in \mathbb{R}^{n\times d}$ (resp. V_R) the matrix formed by the columns $1, \dots, d$ of the matrix $\tilde{V}$ (resp. $\tilde{V}_R$). The matrix $V^* \in \mathbb{R}^{n\times d}$ is the orthonormal matrix with $i-$th column $\frac{1}{\sqrt{n}}\left(Y_{1,1}(X_i), \dots, Y_{1,d}(X_i)\right)$. The matrices G^*, G, G_R and G^*_{proj} are defined as follows

$$G^* := \frac{1}{c_1}V^*(V^*)^\top, \qquad G := \frac{1}{c_1}VV^\top$$
$$G_R := \frac{1}{c_1}V_RV_R^\top, \qquad G^*_{proj} := V^*((V^*)^\top V^*)^{-1}(V^*)^\top.$$

G^*_{proj} is the projection matrix for the columns span of the matrix V^*. Using the triangle inequality we have

$$\|G^* - G\|_F \le \|G^* - G^*_{proj}\|_F + \|G^*_{proj} - G_R\|_F + \|G_R - G\|_F.$$

Step 1: Bounding $\|G - G_R\|_F$. Since the columns of the matrices V and V_R correspond respectively to the eigenvectors of the matrices T_n and $T_{R,n}$, applying the Davis Kahan sinus Theta Theorem (cf. Theorem 3.26) gives that there exists $O \in \mathbb{R}^{d\times d}$ such that

$$\|VO - V_R\|_F \le \frac{2^{3/2}\|T_n - T_{R,n}\|_F}{\Delta},$$

where $\Delta := \min_{k\in\{0,2,3,\dots,R\}} |p_1^* - p_k^*| \ge \Delta^* = \min_{k\in\mathbb{N},\, k\neq 1} |p_1^* - p_k^*|$. Using Lemma 3.25 and $c_1 = \frac{d}{d-2}$, we get that

$$\|G - G_R\|_F = \frac{d-2}{d}\|VO(VO)^\top - V_RV_R^\top\|_F \le 2\|VO - V_R\|_F.$$

Hence, using the proof of Theorem 3.2, we get that with probability at least $1 - 1/n^2$,

$$\|G - G_R\|_F \le 2\|VO - V_R\|_F \le \frac{C}{\Delta^*}\left(\frac{n}{\log^2 n}\right)^{-\frac{s}{2s+d-1}},$$

where $C > 0$ is a constant.

Step 2: Bounding $\|G^* - G^*_{proj}\|_F$. To bound $\|G^* - G^*_{proj}\|_F$, we apply first Lemma 3.27 with $B = V^*$. This leads to

$$\|G^* - G^*_{proj}\|_F \le \|\mathrm{Id}_d - (V^*)^\top V^*\|_F \le \sqrt{d}\|\mathrm{Id}_d - (V^*)^\top V^*\|.$$

Using a proof rigorously analogous to the proof of Lemma 3.21, it holds with probability at least $1 - \gamma$ and for n large enough,

$$\|\mathrm{Id}_d - (V^*)^\top V^*\| \le 4\sqrt{\frac{d\log(e/\gamma)}{n}}.$$

We get by choosing $\gamma = 1/n^2$ that it holds with probability at least $1 - 1/n^2$,

$$\|\mathrm{Id}_d - (V^*)^\top V^*\| \le C'\sqrt{\frac{d\log(n)}{n}},$$

where $C' > 0$ is a universal constant.

Step 3: Bounding $\|G^*_{proj} - G_R\|_F$. We proceed exactly like in Araya and De Castro [2019] but we provide here the proof for completeness. Since G^*_{proj} and G_R are projectors we have, using for example [Bhatia, 1996, p.202],

$$\|G^*_{proj} - G_R\|_F = 2\|G^*_{proj}G_R^\perp\|_F. \tag{3.31}$$

We use Theorem 3.28 with $E = G^*_{proj}$, $F = G_R^\perp$, $B = T_{R,n}$ and $A = T_{R,n} + H$ where

$$H = \tilde{X}_{R,n} K_R \tilde{X}_{R,n}^\top - X_{R,n} K_R X_{R,n},$$

where the columns of the matrix $\tilde{X}_{R,n}$ are obtained using a Gram-Schmidt orthonormalization process on the columns of $X_{R,n}$. Hence there exists a matrix L such that $\tilde{X}_{R,n} = X_{R,n}(L^{-1})^\top$. This matrix L is such that a Cholesky decomposition of $X_{R,n}^\top X_{R,n}$ reads as $LL^\top$.
A and B are symmetric matrices thus we can apply Theorem 3.28. On the event $\mathcal{E}$, we can take $S_1 = (\lambda_1 - \frac{\Delta^*}{8}, \lambda_1 + \frac{\Delta^*}{8})$ and $S_2 = \mathbb{R}\backslash(\lambda_1 - \frac{7\Delta^*}{8}, \lambda_1 + \frac{7\Delta^*}{8})$. By Theorem 3.28 we get

$$\|G^*_{proj} G_R^\perp\|_F \leq \frac{\|A - B\|_F}{\Delta^*} = \frac{\|H\|_F}{\Delta^*}. \tag{3.32}$$

We only need to bound $\|H\|_F$.

$$\begin{aligned} \|H\|_F &\leq \|L^{-\top} K_R L^{-1} - K_R\|_F \|X_{R,n}^\top X_{R,n}\| \\ &\leq \|K_R\|_F \|L^{-1}L^{-\top} - \mathrm{Id}_{\widetilde{R}}\| \|X_{R,n}^\top X_{R,n}\|, \end{aligned} \tag{3.33}$$

where the last inequality comes from Lemma 3.29. From the previous remarks on the matrix L, we directly get

$$\|L^{-1}L^{-\top} - \mathrm{Id}_{\widetilde{R}}\| = \| \left(X_{R,n}^\top X_{R,n}\right)^{-1} - \mathrm{Id}_{\widetilde{R}}\|.$$

Using the notations of the proof of Theorem 3.20 which is provided in Section 3.10.3.1, we get

$$\|L^{-1}L^{-\top} - \mathrm{Id}_{\widetilde{R}}\| \|X_{R,n}^\top X_{R,n}\| = \|X_{R,n}^\top X_{R,n} - \mathrm{Id}_{\widetilde{R}}\| = \|E_{R,n}\|.$$

Noticing further that $\|K_R\|_F^2 \leq \sum_{k\geq 0}(p_k^*)^2 d_k = \|\mathbf{p}\|_2^2 \leq 2$ (because $|\mathbf{p}| \leq 1$), Eq.(3.33) becomes

$$\|H\|_F \leq \sqrt{2}\|E_{R,n}\|. \tag{3.34}$$

Using Lemma 3.21, it holds with probability at least $1 - \gamma$ and for n large enough,

$$\|E_{R,n}\| \leq 4\sqrt{\frac{\widetilde{R}}{n}\ln(2/\gamma)}. \tag{3.35}$$

Since $\widetilde{R} = \mathcal{O}\left(R^{d-1}\right)$ and $R = \mathcal{O}\left(\left(n/\log^2 n\right)^{\frac{1}{2s+d-1}}\right)$, we obtain using Eqs.(3.31), (3.32), (3.34) and (3.35) that with probability at least $1 - 1/n^2$ it holds

$$\|G^*_{proj} - G_R\|_F = 2\|G^*_{proj} G_R^\perp\|_F \leq \frac{C_d}{\Delta^*}\left(\frac{n}{\log^2(n)}\right)^{\frac{-s}{2s+d-1}},$$

where $C_d > 0$ is a constant that may depend on d and on constants related to the Markov chain $(X_i)_{i\geq 1}$.

Conclusion. We proved that on the event $\mathcal{E}$, it holds with probability at least $1 - 3/n^2$,

$$\|G^* - G\|_F \leq D_1\left(\frac{n}{\log^2(n)}\right)^{\frac{-s}{2s+d-1}},$$

where $D_1 > 0$ is a constant that depends on Δ^*, d and on constants related to the Markov chain $(X_i)_{i\geq 1}$. Moreover, Eq.(3.39) from the proof of Proposition 3.24 gives that on the event $\mathcal{E}$, we have

$$\|G - \hat{G}\|_F = \frac{d-2}{d}\|VV^\top - \hat{V}\hat{V}^\top\|_F \leq \frac{2^{\frac{9}{2}}\sqrt{d}\|T_n - \widehat{T}_n\|}{3\Delta^*}.$$

Using the concentration result from Bandeira and van Handel [2016] on spectral norm of centered random matrix with independent entries we get that there exists some constant $D_2 > 0$ such that with

probability at least $1 - 1/n^2$ it holds

$$\|G - \hat{G}\|_F \leq D_2 \frac{\sqrt{\log n}}{n}.$$

Using again Proposition 3.24, we know that for n large enough, $\mathbb{P}(\mathcal{E}) \geq 1 - 1/n^2$. We conclude that for n large enough, it holds with probability at least $1 - 5/n^2$,

$$\|G^* - \hat{G}\|_F \leq D_3 \left(\frac{n}{\log^2(n)}\right)^{\frac{-s}{2s+d-1}},$$

for some constant $D_3 > 0$ that depends on Δ^*, d and on constants related to the Markov chain $(X_i)_{i\geq 1}$ (see Theorem 3.20 for details).

3.10.4.1 Proof of Proposition 3.24

First part of the proof Let us consider $\gamma > 0$.
Using the concentration of spectral norm for random matrices with independent entries from Bandeira and van Handel [2016], there exists a universal constant C_0 such that

$$\mathbb{P}\left(\|T_n - \widehat{T}_n\| \leq \frac{3\sqrt{2D_0}}{n} + C_0 \frac{\sqrt{\log n/\gamma}}{n}\right) \leq \gamma,$$

where denoting $Y = T_n - \widehat{T}_n$, we define $D_0 := \max_{1\leq i\leq n} \sum_{j=1}^n Y_{i,j}\left(1 - Y_{i,j}\right)$. We deduce that for n large enough, it holds with probability at least $1 - \gamma/4$,

$$\|T_n - \widehat{T}_n\| \leq \frac{(\Delta^*)^2}{2^{\frac{13}{2}}\sqrt{d}}. \tag{3.36}$$

Using now Theorem 3.2, it holds with probability at least $1 - \gamma/4$ for n large enough

$$\delta_2\left(\lambda(T_n), \lambda(\mathbb{T}_W)\right) \leq C\left(\frac{\log^2 n}{n}\right)^{\frac{s}{2s+d-1}} \leq \frac{\Delta^*}{8}. \tag{3.37}$$

Putting together Eq.(3.36) and Eq.(3.37), we deduce that for n large enough,

$$\mathbb{P}\left(\mathcal{E}\right) \geq 1 - \gamma/2.$$

Second part of the proof In the following, we work on the event $\mathcal{E}$. Since $\Delta^* > 0$ by assumption, we get that $p_1^* = \lambda_1^* = \cdots = \lambda_d^*$ is the only eigenvalue of $\mathbb{T}_W$ with multiplicity d. Indeed, all eigenvalue p_k^* with $k > d$ has multiplicity $d_k > d$ and p_0^* has multiplicity 1. Moreover, from Eq.(3.37), we have that there exists a unique set of d eigenvalues of T_n, denoted $\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_d}$, such that they are at a distance least $3\Delta^*/4$ away from the other eigenvalues, i.e.

$$\Delta := \min_{\nu_1 \in \lambda(T_n)\setminus\{\lambda_{i_1},\lambda_{i_2},\ldots,\lambda_{i_d}\}} \max_{\nu_2\in\{\lambda_{i_1},\lambda_{i_2},\ldots,\lambda_{i_d}\}} |\nu_1 - \nu_2| \geq \frac{3\Delta^*}{4}. \tag{3.38}$$

Let us form the matrix $V \in \mathbb{R}^{n\times d}$ where the k-th column is the eigenvector of T_n associated with the eigenvalue λ_{i_k}. We denote further $G := VV^\top/d$. Let $\hat{V} \in \mathbb{R}^{n\times d}$ be the matrix with columns corresponding to the eigenvectors associated to eigenvalues $\hat{\lambda}_{i_1}, \hat{\lambda}_{i_2}, \ldots, \hat{\lambda}_{i_d}$ of $\widehat{T}_n$ and $\hat{G} := \hat{V}\hat{V}^\top/d$. Using Theorem 3.26 there exists some orthonormal matrix $O \in \mathbb{R}^{d\times d}$ such that

$$\|VO - \hat{V}\|_F \leq \frac{2^{\frac{3}{2}} \min\{\sqrt{d}\|T_n - \widehat{T}_n\|, \|T_n - \widehat{T}_n\|_F\}}{\Delta}.$$

Denoting $\lambda_{i_1}^{sort} \geq \lambda_{i_2}^{sort} \geq \cdots \geq \lambda_{i_d}^{sort}$ (resp. $\hat{\lambda}_{i_1}^{sort} \geq \hat{\lambda}_{i_2}^{sort} \geq \cdots \geq \hat{\lambda}_{i_d}^{sort}$) the sorted version of the eigenvalues $\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_d}$ (resp. $\hat{\lambda}_{i_1}, \hat{\lambda}_{i_2}, \ldots, \hat{\lambda}_{i_d}$), we have

$$\begin{aligned}
&\left[\sum_{k=1}^{d}\left(\lambda_{i_k}^{sort} - \hat{\lambda}_{i_k}^{sort}\right)^2\right]^{1/2} \\
\leq\ & \|VV^\top - \hat{V}\hat{V}^\top\|_F \quad \text{(Hoffman-Wielandt inequality [Bhatia, 1996, Thm VI.4.1])} \\
\leq\ & 2\|VO - \hat{V}\|_F \qquad \text{(using Lemma 3.25)} \\
\leq\ & \frac{2^{\frac{5}{2}}\min\{\sqrt{d}\|T_n - \widehat{T}_n\|, \|T_n - \widehat{T}_n\|_F\}}{\Delta} \\
\leq\ & \frac{2^{\frac{9}{2}}\min\{\sqrt{d}\|T_n - \widehat{T}_n\|, \|T_n - \widehat{T}_n\|_F\}}{3\Delta^*} \quad \text{(using Eq.(3.38))} \qquad (3.39) \\
\leq\ & \Delta^*/8. \quad \text{(using Eq.(3.36))}
\end{aligned}$$

Using the triangle inequality, we get that

$$\hat{\Delta} := \min_{\nu_1 \in \lambda(\widehat{T}_n)\setminus\{\hat{\lambda}_{i_1}, \hat{\lambda}_{i_2}, \ldots, \hat{\lambda}_{i_d}\}} \max_{\nu_2 \in \{\hat{\lambda}_{i_1}, \hat{\lambda}_{i_2}, \ldots, \hat{\lambda}_{i_d}\}} |\nu_1 - \nu_2| \geq \frac{\Delta^*}{2}. \qquad (3.40)$$

We proved that on the event $\mathcal{E}$, the eigenvalues in $\Lambda_1 := \{\hat{\lambda}_{i_1}, \ldots, \hat{\lambda}_{i_d}\}$ are at distance at least $\Delta^*/2$ from the other eigenvalues of $\widehat{T}_n$ (cf. Eq.(3.40)) and are at distance at most $\Delta^*/8$ of the eigenvalues $\lambda_{i_1}, \ldots, \lambda_{i_d}$ of T_n. We could have done this analysis for different eigenvalues. Let us consider some $k \geq 0$. Eq.(3.37) shows that on the event $\mathcal{E}$, there exists a set of d_k eigenvalues of T_n which concentrate around p_k^* and such that it has diameter at most $\Delta^*/4$. Weyl's inequality (cf. [Bhatia, 1996, p.63]) proves that there exist d_k eigenvalues of $\widehat{T}_n$ that are at distance at most $\Delta^*/4$ from p_k^*. If we consider now a subset $L \neq \Lambda_1$ of d eigenvalues of $\widehat{T}_n$, then the previous analysis shows that there exists some eigenvalue $\hat{\lambda}$ of $\widehat{T}_n$ which is not in L and that is at distance at most $\Delta^*/4$ from one eigenvalue in L. Using Eq.(3.38), we deduce that Algorithm (HEiC) returns $\hat{G} = \hat{V}\hat{V}^\top/d$ where the columns of $\hat{V}$ correspond to the eigenvectors of $\widehat{T}_n$ associated to the eigenvalues in Λ_1.

3.10.4.2 Useful results

Lemma 3.25. *Let A, B be two matrices in $\mathbb{R}^{n\times d}$ then*

$$\|AA^\top - BB^\top\|_F \leq (\|A\| + \|B\|)\|A - B\|_F.$$

If $A^\top A = B^\top B = \mathrm{Id}$ then

$$\|AA^\top - BB^\top\|_F \leq 2\|A - B\|_F.$$

Proof of Lemma 3.25.

$$\begin{aligned}
\|AA^\top - BB^\top\|_F &= \|(A-B)A^\top + B(A^\top - B^\top)\|_F \\
&\leq \|A(A-B)^\top\|_F + \|(B-A)B^\top\|_F \\
&\leq \|(A\otimes \mathrm{Id}_n)vec(A-B)\|_2 + \|(\mathrm{Id}_d \otimes B)vec(A-B)^\top\|_2 \\
&\leq (\|A\otimes \mathrm{Id}_n\| + \|\mathrm{Id}_d \otimes B\|)\,\|A-B\|_F \\
&= (\|A\| + \|B\|)\|A-B\|_F,
\end{aligned}$$

where $vec(\cdot)$ represents the vectorization of a matrix that is its transformation into a column vector and $\otimes$ is the notation for the Kronecker product between two matrices. □

Theorem 3.26. *(Davis-Kahan Theorem, cf. [Yu et al., 2014]) Let Σ and $\hat{\Sigma}$ be two symmetric $\mathbb{R}^{n\times n}$ matrices with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$ respectively. For $1 \leq r \leq s \leq n$ fixed, we assume that $\min\{\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}\} > 0$ where $\lambda_0 := \infty$ and $\lambda_{n+1} = -\infty$. Let $d = s - r + 1$ and V and $\hat{V}$ two matrices in $\mathbb{R}^{n\times d}$ with columns $(v_r, v_{r+1}, \ldots, v_s)$ and $(\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s)$ respectively, such that $\Sigma v_j = \lambda_j v_j$ and*

$\hat{\Sigma}\hat{v}_j = \lambda_j \hat{v}_j$. *Then there exists an orthogonal matrix* $\hat{O}$ *in* $\mathbb{R}^{d\times d}$ *such that*

$$\|\hat{V}\hat{O} - V\|_F \leq \frac{2^{3/2}\min\{\sqrt{d}\|\Sigma - \hat{\Sigma}\|, \|\Sigma - \hat{\Sigma}\|_F\}}{\min\{\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}\}}.$$

Lemma 3.27. *Let* B *be a* $n \times d$ *matrix with full column rank. Then we have*

$$\|BB^\top - B(B^\top B)^{-1}B^\top\|_F = \|\mathrm{Id}_d - B^\top B\|_F.$$

Proof of Lemma 3.27. Using the cyclic property of the trace, we have

$$\begin{aligned}
&\|BB^\top - B(B^\top B)^{-1}B^\top\|_F^2 \\
&= \|B\left(\mathrm{Id}_d - (B^\top B)^{-1}\right)B^\top\|_F^2 \\
&= \mathrm{Tr}\left(B\left(\mathrm{Id}_d - (B^\top B)^{-1}\right)B^\top B\left(\mathrm{Id}_d - (B^\top B)^{-1}\right)B^\top\right) \\
&= \mathrm{Tr}\left(B^\top B\left(\mathrm{Id}_d - (B^\top B)^{-1}\right)B^\top B\left(\mathrm{Id}_d - (B^\top B)^{-1}\right)\right) \\
&= \mathrm{Tr}\left(\left(B^\top B - \mathrm{Id}_d\right)\left(B^\top B - \mathrm{Id}_d\right)\right) \\
&= \|\mathrm{Id}_d - B^\top B\|_F^2.
\end{aligned}$$

□

Theorem 3.28. *(cf. [Bhatia, 1996, ThmVII.3.4]) Let* A *and* B *be two normal operators and* S_1 *and* S_2 *two sets separated by a strip of size* δ. *Let* E *be the orthogonal projection matrix of the eigenspaces of* A *with eigenvalues inside* S_1 *and* F *be the orthogonal projection matrix of the eigenspaces of* B *with eigenvalues inside* S_2. *Then*

$$\|EF\|_F \leq \frac{1}{\delta}\|E(A-B)F\|_F \leq \frac{1}{\delta}\|A - B\|_F.$$

Lemma 3.29. *(Ostrowski's inequality) Let* $A \in \mathbb{R}^{n\times n}$ *be a Hermitian matrix and* $S \in \mathbb{R}^{d\times n}$ *be a general matrix then*

$$\|SAS^\top - A\|_F \leq \|A\|_F \times \|S^\top S - \mathrm{Id}_n\|.$$

3.10.5 Proof of Proposition 3.12

Notice that for any $i \in [n]$,

$$\mathbb{P}\left(g_i(\mathbf{D}_{1:n}) \neq A_{i,n+1}\right) = \mathbb{E}\left[\mathbb{1}_{g_i(\mathbf{D}_{1:n})\neq A_{i,n+1}}\right] = \mathbb{E}\mathbb{E}\left[\mathbb{1}_{g_i(\mathbf{D}_{1:n})\neq A_{i,n+1}} \mid \mathbf{D}_{1:n}\right],$$

and that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{g_i(\mathbf{D}_{1:n})\neq A_{i,n+1}} \mid \mathbf{D}_{1:n}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{g_i(\mathbf{D}_{1:n})=1}\mathbb{1}_{A_{i,n+1}=0} \mid \mathbf{D}_{1:n}\right] + \mathbb{E}\left[\mathbb{1}_{g_i(\mathbf{D}_{1:n})=0}\mathbb{1}_{A_{i,n+1}=1} \mid \mathbf{D}_{1:n}\right] \\
&= \eta_i(\mathbf{D}_{1:n})\mathbb{1}_{g_i(\mathbf{D}_{1:n})=0} + (1-\eta_i(\mathbf{D}_{1:n}))\,\mathbb{1}_{g_i(\mathbf{D}_{1:n})=1},
\end{aligned}$$

which leads to

$$\mathbb{P}\left(g_i(\mathbf{D}_{1:n}) \neq A_{i,n+1}\right) = \mathbb{E}\left[\eta_i(\mathbf{D}_{1:n})\mathbb{1}_{g_i(\mathbf{D}_{1:n})=0} + (1-\eta_i(\mathbf{D}_{1:n}))\,\mathbb{1}_{g_i(\mathbf{D}_{1:n})=1}\right].$$

By definition of the Bayes classifier g^*, we have for any $i \in [n]$,

$$\begin{aligned}
&\mathbb{P}\left(g_i^*(\mathbf{D}_{1:n}) \neq A_{i,n+1}\right) \\
&= \mathbb{E}\left[\eta_i(\mathbf{D}_{1:n})\mathbb{1}_{\eta_i(\mathbf{D}_{1:n})<\frac{1}{2}} + (1-\eta_i(\mathbf{D}_{1:n}))\,\mathbb{1}_{\eta_i(\mathbf{D}_{1:n})\geq\frac{1}{2}}\right] \\
&= \mathbb{E}\left[\min\left\{\eta_i(\mathbf{D}_{1:n}), 1-\eta_i(\mathbf{D}_{1:n})\right\}\left(\mathbb{1}_{\eta_i(\mathbf{D}_{1:n})\geq\frac{1}{2}} + \mathbb{1}_{\eta_i(\mathbf{D}_{1:n})<\frac{1}{2}}\right)\right] \\
&= \mathbb{E}\left[\min\left\{\eta_i(\mathbf{D}_{1:n}), 1-\eta_i(\mathbf{D}_{1:n})\right\}\right]
\end{aligned}$$

Given another classifier g, we have for any $i \in [n]$,

$$\begin{aligned}
& \mathbb{P}\left(g_i(\mathbf{D}_{1:n}) \neq A_{i,n+1}\right) - \mathbb{P}\left(g_i^*(\mathbf{D}_{1:n}) \neq A_{i,n+1}\right) \\
= \; & \mathbb{E}\Bigg[\eta_i(\mathbf{D}_{1:n})\mathbb{1}_{g_i(\mathbf{D}_{1:n})=0} + (1-\eta_i(\mathbf{D}_{1:n}))\,\mathbb{1}_{g_i(\mathbf{D}_{1:n})=1} \\
& \qquad - \left(\eta_i(\mathbf{D}_{1:n})\mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=0} + (1-\eta_i(\mathbf{D}_{1:n}))\,\mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=1}\right)\Bigg] \\
= \; & \mathbb{E}\big[\eta_i(\mathbf{D}_{1:n})\left(\mathbb{1}_{g_i(\mathbf{D}_{1:n})=0} - \mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=0}\right) \\
& \quad + (1-\eta_i(\mathbf{D}_{1:n}))\left(\mathbb{1}_{g_i(\mathbf{D}_{1:n})=1} - \mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=1}\right)\big] \\
= \; & \mathbb{E}\left[(2\eta_i(\mathbf{D}_{1:n})-1)\left(\mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=1} - \mathbb{1}_{g_i(\mathbf{D}_{1:n})=1}\right)\right],
\end{aligned}$$

where we used that $g(\mathbf{D}_{1:n})$ takes only the values 0 and 1, so that

$$\mathbb{1}_{g_i(\mathbf{D}_{1:n})=0} - \mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=0} = \left(\mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=1} - \mathbb{1}_{g_i(\mathbf{D}_{1:n})=1}\right).$$

Since

$$\begin{aligned}
\mathbb{1}_{g_i^*(\mathbf{D}_{1:n})=1} - \mathbb{1}_{g_i(\mathbf{D}_{1:n})=1} &= \begin{cases} 1 & \text{if } g_i^*(\mathbf{D}_{1:n}) = 1 \text{ and } g_i(\mathbf{D}_{1:n}) = 0 \\ 0 & \text{if } g_i^*(\mathbf{D}_{1:n}) = g_i(\mathbf{D}_{1:n}) \\ -1 & \text{if } g_i^*(\mathbf{D}_{1:n}) = 0 \text{ and } g_i(\mathbf{D}_{1:n}) = 1 \end{cases} \\
&= \mathbb{1}_{g_i^*(\mathbf{D}_{1:n}) \neq g_i(\mathbf{D}_{1:n})}\ \mathrm{sgn}(\eta_i(\mathbf{D}_{1:n}) - 1/2),
\end{aligned}$$

we deduce that

$$\begin{aligned}
& \mathbb{P}\left(g_i(\mathbf{D}_{1:n}) \neq A_{i,n+1}\right) - \mathbb{P}\left(g_i^*(\mathbf{D}_{1:n}) \neq A_{i,n+1}\right) \\
& = 2\mathbb{E}\left[\left|\eta_i(\mathbf{D}_{1:n}) - \frac{1}{2}\right| \times \mathbb{1}_{g_i(\mathbf{D}_{1:n}) \neq g_i^*(\mathbf{D}_{1:n})}\right],
\end{aligned}$$

which concludes the proof.

Chapter 4

Concentration inequality for U-statistics

Chapter Abstract

In this chapter, we prove a new concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. Contrary to previous works, we consider general state space and index-dependent kernel functions that are not assumed to be symmetric or smooth. We provide examples of Markov chains satisfying our conditions and we stress the importance for learning theory to work with index-dependent kernels. We give first an Hoeffding-type bound that holds without any condition on the initial distribution of the chain and we provide a Bernstein-type concentration result when the chain is stationary.

Chapter Content

4.1 Introduction

Concentration of measure has been intensely studied during the last decades since it finds application in large span of topics such as model selection (see Massart [2007] and Lerasle et al. [2016]), statistical learning (see Clémençon et al. [2020]), online learning (see Wang et al. [2012]) or random graphs (see Chapter 3). Important contributions in this field are those concerning U-statistics. A U-statistic of order m is a sum of the form

$$\sum_{1\leq i_1<\cdots<i_m\leq n} h_{i_1,\ldots,i_m}(X_{i_1},\ldots,X_{i_m}),$$

where $X_1,\ldots,X_n$ are independent random variables taking values in a measurable space (E,Σ) (with E Polish) and with respective laws P_i and where $h_{i_1,\ldots,i_m}$ are measurable functions of m variables $h_{i_1,\ldots,i_m}: E^m\to\mathbb{R}$.

One important exponential inequality for U-statistics was provided by Arcones and Giné [1993] using a Rademacher chaos approach. Their result holds for bounded and canonical (or degenerate) kernels, namely satisfying for all $i_1,\ldots,i_m\in[n]:=\{1,\ldots,n\}$ with $i_1<\cdots<i_m$ and for all $x_1,\ldots,x_m\in E$,

$$\left\|h_{i_1,\ldots,i_m}\right\|_\infty<\infty \quad\text{and}\quad \forall j\in[1,n]\,,\ \mathbb{E}_{X_j}\left[h_{i_1,\ldots,i_m}(x_1,\ldots,x_{j-1},X_j,x_{j+1},\ldots,x_m)\right]=0\,.$$

They proved that in the degenerate case, the convergence rates for U-statistics are expected to be $n^{m/2}$. Relying on precise moment inequalities of Giné et al. [2000] improved the result from Arcones and Giné [1993] by providing the optimal four regimes of the tail, namely Gaussian, exponential, Weibull of orders $2/3$ and $1/2$. In the specific case of order 2 U-statistics, Houdré and Reynaud-Bouret [2003] recovered the result from Giné et al. [2000] by replacing the moment estimates by martingales type inequalities, giving as a by-product explicit constants. When the kernels are unbounded, it was shown that some results can be extended provided that the random variables $h_{i_1,\ldots,i_m}(X_{i_1},\ldots,X_{i_m})$ have sufficiently light tails. One can mention [Eichelsbacher and Schmock, 2003, Theorem 3.26] where an exponential inequality for U-statistics with a single Banach-space valued, unbounded and canonical kernel is proved. Their approach is based on a decoupling argument originally obtained by de la Peña and Montgomery-Smith [1995] and the tail behavior of the summands is controlled by assuming that the kernel satisfies the so-called weak Cramér condition. It is now well-known that with heavy-tailed distribution for $h_{i_1,\ldots,i_m}(X_{i_1},\ldots,X_{i_m})$ we cannot expect to get exponential inequalities anymore. Nevertheless working with kernels that have finite p-th moment for some $p\in(1,2]$, Joly and Lugosi in Joly and Lugosi [2016] construct an estimator of the mean of the U-process using the median-of-means technique that performs as well as the classical U-statistic with bounded kernels.

All the above mentioned results consider that the random variables $(X_i)_{i\geq1}$ are independent. This condition can be prohibitive for practical applications since modelization of real phenomena often involves some dependence structure. The simplest and the most widely used tool to incorporate such dependence is Markov chain. One can give the example of Reinforcement Learning (see Sutton and Barto [2018]) or Biology (see Suchard et al. [2001]). Recent works provide extensions of the classical concentration results to the Markovian settings as Adamczak [2008], Clémençon et al. [2020], Fan et al. [2021], Jiang et al. [2018], Paulin [2015]. The asymptotic behaviour of U-statistics in the Markovian setup has already been investigated by several papers. We refer to Bertail and Clémençon [2011] where the authors proved a Strong Law of Large Numbers and a Central Limit Theorem proved for U-statistics of order 2 using the *renewal approach* based on the splitting technique. One can also mention Eichelsbacher and Schmock [2001] regarding large deviation principles. However, there are only few results for the non-asymptotic behaviour of tails of U-statistics in a dependent framework. The first results were provided in Borisov and Volodko [2015] and Han [2018] where exponential inequalities for U-statistics of order $m\geq2$ of time series under mixing conditions are proved. Those works were improved by Shen et al. [2020] where a Hoeffding-type inequality for V and U-statistics is provided under conditions on the time dependent process that are easier to check in practice. In Section 4.3.3.3, we describe in details the result of Shen et al. [2020] and the differences with our work. Let us point out that all the above mentioned works regarding non-asymptotic tail bound for U-statistics in a dependent framework consider a fixed kernel, namely $h\equiv h_{i_1,\ldots,i_m}$ for all $i_1,\ldots,i_m$. Our work is the first to consider time dependent kernel functions which makes the theoretical analysis more challenging since the standard splitting method can be unworkable (cf. Section 4.2.5). In Section 4.3.3.4 and 4.3.3.5, we stress the importance of working with index-dependent kernels for practical applications and show on specific examples that one can reach significantly faster convergence rates with this approach.

For the first time, we provide in this thesis a Bernstein-type concentration inequality for U-statistics of order 2 in a dependent framework with kernels that may depend on the indexes of the sum and that are not assumed to be symmetric or smooth. We work on a general state space with bounded kernels that are π-canonical. This latter notion was first introduced in Fort et al. [2012] who proved a variance inequality for U-statistics of ergodic Markov chains. Our Bernstein bound holds for stationary chains but we provide a Hoeffding-type inequality without any assumption on the initial distribution of the Markov chain.

4.1.1 Outline

In Section 4.2, we present and comment the assumptions under which our main results hold. In Section 4.3.1, we define and comment the key quantities involved in our results and we present our exponential inequalities with Theorems 4.3 and 4.4 in Section 4.3.2. Section 4.3.3 is dedicated to discussions where we give examples of Markov chains satisfying our assumptions and where we compare our results with the independent case. The proofs of both Theorems are presented in Section 4.4. In Section 4.5, we provide the proofs of some technical lemmas.

4.2 Assumptions and Notations

We consider a Markov chain $(X_i)_{i\geq 1}$ with transition kernel $P : E\times E \to \mathbb{R}$ taking values in a measurable space (E, Σ), and we introduce bounded functions $h_{i,j} : E^2 \to \mathbb{R}$. In this section, we describe the different assumptions on the Markov chain $(X_i)_{i\geq 1}$ and on the functions $h_{i,j}$ that we will consider in Theorems 4.3 and 4.4 presented in the next section.

4.2.1 Uniform ergodicity

Assumption 1. *The Markov chain $(X_i)_{i\geq 1}$ is ψ-irreducible for some maximal irreducibility measure ψ on Σ (see [Meyn and Tweedie, 1993, Section 4.2]). Moreover, there exist an integer $m \geq 1$, $\delta_m > 0$ and some probability measure μ such that*

$$\forall x \in E,\ \forall A \in \Sigma, \quad \delta_m \mu(A) \leq P^m(x, A).$$

We denote by π the unique stationary distribution of the Markov chain $(X_i)_{i\geq 1}$.

For the reader familiar with the theory of Markov chains, Assumption 1 states that the whole space E is a small set which is equivalent to the uniform ergodicity of the Markov chain $(X_i)_{i\geq 1}$ (see [Meyn and Tweedie, 1993, Theorem 16.0.2]), namely there exist constants $0 < \rho < 1$ and $L > 0$ such that

$$\|P^n(x, \cdot) - \pi\|_{\mathrm{TV}} \leq L\rho^n, \qquad \forall n \geq 0,\ \pi\text{−a.e } x \in E,$$

where π is the unique stationary distribution of the chain $(X_i)_{i\geq 1}$ and where for any measure ω on (E, Σ), $\|\omega\|_{\mathrm{TV}} := \sup_{A\in\Sigma} |\omega(A)|$ is the total variation norm of ω. From [Ferré et al., 2012, section 2.3]), we also know that the Markov chain $(X_i)_{i\geq 1}$ admits an absolute spectral gap $1 - \lambda > 0$ with $\lambda \in [0, 1)$ (thanks to uniform ergodicity). We refer to A.3 or to [Fan et al., 2021, Section 3.1] for a reminder on the spectral gap of Markov chains.

4.2.2 Upper-bounded Markov kernel

Assumption 2 can be read as a reverse Doeblin's condition and allows us to achieve a change of measure in expectations in our proof to work with i.i.d. random variables with distribution ν. As a result, Assumption 2 is the cornerstone of our approach since it allows to decouple the U-statistic in the proof.

Assumption 2. *There exist $\delta_M > 0$ and some probability measure ν such that*

$$\forall x \in E,\ \forall A \in \Sigma, \quad P(x, A) \leq \delta_M \nu(A).$$

Assumption 2 has already been used in the literature (see [Lindsten et al., 2015, Section 4.2]) and was introduced in Del Moral and Guionnet [1999]. This condition can typically require the state space to be compact as highlighted in Lindsten et al. [2015].

Let us describe another situation where Assumption 2 holds. Consider that $(E, \|\cdot\|)$ is a normed space and that for all $x \in E$, $P(x, dy)$ has density $p(x, \cdot)$ with respect to some measure η on (E, Σ). We further assume that there exists an integrable function $u : E \to \mathbb{R}_+$ such that $\forall x, y \in E, \quad p(x, y) \leq u(y)$. Then considering for ν the probability measure with density $u/\|u\|_1$ with respect to η and $\delta_M = \|u\|_1$, Assumption 2 holds.

4.2.3 Exponential integrability of the regeneration time

We introduce some additional notations which will be useful to apply Talagrand concentration result from Samson [2000]. Note that this section is inspired from Adamczak [2008] and [Meyn and Tweedie, 1993, Theorem 17.3.1]. We assume that Assumption 1 is satisfied and we extend the Markov chain $(X_i)_{i\geq 1}$ to a new (so called *split*) chain $(\widetilde{X}_n, R_n) \in E \times \{0, 1\}$ (see [Meyn and Tweedie, 1993, Section 5.1] for a reminder on the splitting technique), satisfying the following properties.

- $(\widetilde{X}_n)_n$ is again a Markov chain with transition kernel P with the same initial distribution as $(X_n)_n$. We recall that π is the stationary distribution on the E.
- if we define $T_1 = \inf\{n > 0 : R_{nm} = 1\}$,

$$T_{i+1} = \inf\{n > 0 : R_{(T_1+\cdots+T_i+n)m} = 1\},$$

  then $T_1, T_2, \ldots$ are well defined and independent. Moreover $T_2, T_3, \ldots$ are i.i.d.
- if we define $S_i = T_1 + \cdots + T_i$, then the "blocks"

$$Y_0 = (\widetilde{X}_1, \ldots, \widetilde{X}_{mT_1+m-1}), \quad \text{and} \quad Y_i = (\widetilde{X}_{m(S_i+1)}, \ldots, \widetilde{X}_{m(S_{i+1}+1)-1}),\ i > 0,$$

  form a one-dependent sequence (i.e. for all i, $\sigma((Y_j)_{j<i})$ and $\sigma((Y_j)_{j>i})$ are independent). Moreover, the sequence $Y_1, Y_2, \ldots$ is stationary and if $m = 1$ the variables $Y_0, Y_1, \ldots$ are independent. In consequence, for any measurable space $(S, \mathcal{B})$ and measurable functions $f : S \to \mathbb{R}$, the variables

$$Z_i = Z_i(f) = \sum_{j=m(S_i+1)}^{m(S_{i+1}+1)-1} f(\widetilde{X}_j), \quad i \geq 1,$$

  constitute a one-dependent sequence (an i.i.d. sequence if $m = 1$). Additionally, if f is π-integrable (recall that π is the unique stationary measure for the chain), then

$$\mathbb{E}[Z_i] = \delta_m^{-1} m \int f d\pi.$$

- the distribution of T_1 depends only on π, P, δ_m, μ, whereas the law of T_2 only on P, δ_m and μ.

Remark. Let us highlight that $(\widetilde{X}_n)_n$ is a Markov chain with transition kernel P and same initial distribution as $(X_n)_n$. Hence for our purposes of estimating the tail probabilities, we will identify $(X_n)_n$ and $(\widetilde{X}_n)_n$.
To derive a concentration inequality, we use the exponential integrability of the regeneration times which is ensured if the chain is uniformly ergodic as stated by Proposition 4.2. A proof can be found in Section 4.5.6.

Definition 4.1. For $\alpha > 0$, define the function $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ with the formula $\psi_\alpha(x) = \exp(x\alpha) - 1$. Then for a random variable X, the α-Orlicz norm is given by

$$\|X\|_{\psi_\alpha} = \inf\left\{\gamma > 0 : \mathbb{E}[\psi_\alpha(|X|/\gamma)] \leq 1\right\}.$$

Proposition 4.2. *If Assumption 1 holds, then*

$$\|T_1\|_{\psi_1} < \infty \quad \textit{and} \quad \|T_2\|_{\psi_1} < \infty, \tag{4.1}$$

where $\|\cdot\|_{\psi_1}$ is the 1-Orlicz norm introduced in Definition 4.1. We denote $\tau := \max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1})$.

4.2.4 π–canonical and bounded kernels

With Assumption 3, we introduce the notion of π-*canonical* kernel which is the counterpart of the canonical property from Giné and Nickl [2016].

Assumption 3. *Let us denote $\mathcal{B}(\mathbb{R})$ the Borel algebra on $\mathbb{R}$. For all $i, j \in [n]$, we assume that $h_{i,j} : (E^2, \Sigma \otimes \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable and is π-canonical, namely*

$$\forall x, y \in E, \quad \mathbb{E}_\pi[h_{i,j}(X, x)] = \mathbb{E}_\pi[h_{i,j}(X, y)] = \mathbb{E}_\pi[h_{i,j}(x, X)] = \mathbb{E}_\pi[h_{i,j}(y, X)].$$

This common expectation will be denoted by $\mathbb{E}_\pi[h_{i,j}]$. Moreover, we assume that for all $i, j \in [n]$, $\|h_{i,j}\|_\infty < \infty$.

Remarks.
• A large span of kernels are π-canonical. This is the case of translation-invariant kernels which have been widely studied in the Machine Learning community. Another example of π-canonical kernel is a rotation invariant kernel when $E = \mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ with π also rotation invariant (see De Castro et al. [2019] or Duchemin and De Castro [2022]).

• The notion of π-canonical kernels is the counterpart of canonical kernels in the i.i.d. framework (see for example Houdré and Reynaud-Bouret [2003]). Note that we are not the first to introduce the notion of π-canonical kernels working with Markov chains. In Fort et al. [2012], the authors provide a variance inequality for U-statistics whose underlying sequence of random variables is an ergodic Markov Chain. Their results holds for π-canonical kernels as stated with [Fort et al., 2012, Assumption A2].

• Note that if the kernels $h_{i,j}$ are not π-canonical, the U-statistic decomposes into a linear term and a π-canonical U-statistic. This is called the *Hoeffding decomposition* (see [Giné and Nickl, 2016, p.176]) and takes the following form

$$\begin{aligned}\sum_{i\neq j}\left(h_{i,j}(X_i, X_j) - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}[h_{i,j}(X,Y)]\right) &= \sum_{i\neq j}\widetilde{h}_{i,j}(X_i, X_j) - \mathbb{E}_\pi\left[\widetilde{h}_{i,j}\right] \\ &+ \sum_{i\neq j}\left(\mathbb{E}_{X\sim\pi}\left[h_{i,j}(X, X_j)\right] - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}\left[h_{i,j}(X,Y)\right]\right) \\ &+ \sum_{i\neq j}\left(\mathbb{E}_{X\sim\pi}\left[h_{i,j}(X_i, X)\right] - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}\left[h_{i,j}(X,Y)\right]\right),\end{aligned}$$

where for all j, the kernel $\widetilde{h}_{i,j}$ is π-canonical with

$$\forall x, y \in E, \quad \widetilde{h}_{i,j}(x, y) = h_{i,j}(x, y) - \mathbb{E}_{X\sim\pi}\left[h_{i,j}(x, X)\right] - \mathbb{E}_{X\sim\pi}\left[h_{i,j}(X, y)\right].$$

4.2.5 Additional technical assumption

In the case where the kernels $h_{i,j}$ depend on both i and j, we need Assumption 4.(ii) to prove Theorem 4.3. Assumption 4.(ii) is a mild condition on the initial distribution of the Markov chain that is used when we apply Bernstein's inequality for Markov chains from Proposition 4.19.

Assumption 4. *At least one of the following conditions holds.*
(i) *For all $i, j \in [n]$, $h_{i,j} \equiv h_{1,j}$, i.e. the kernel function $h_{i,j}$ does not depend on i.*
(ii) *The initial distribution of the Markov chain $(X_i)_{i\geq 1}$, denoted χ, is absolutely continuous with respect to the stationary measure π and its density $\frac{d\chi}{d\pi}$ has finite p-moment for some $p \in (1, \infty]$, i.e*

$$\infty > \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} := \begin{cases} \left[\int \left|\frac{d\chi}{d\pi}\right|^p d\pi\right]^{1/p} & \text{if } p < \infty, \\ \operatorname{ess\,sup} \left|\frac{d\chi}{d\pi}\right| & \text{if } p = \infty. \end{cases}$$

In the following, we will denote $q = \frac{p}{p-1} \in [1, \infty)$ (with $q = 1$ if $p = +\infty$) which satisfies $\frac{1}{p} + \frac{1}{q} = 1$.

Assumption 4 is needed at one specific step of our proof where we need to bound with high probability

$$\sum_{j=2}^{n} \mathbb{E}\Big[\Big|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\Big|^k\Big], \quad \text{with} \quad \forall i, j, \quad \forall x, y \in E, \quad p_{i,j}(x, y) := h_{i,j}(x, y) - \mathbb{E}_\pi[h_{i,j}],$$

and where $(X'_j)_{j\geq 1}$ are i.i.d. random variables with distribution ν from Assumption 2. In the case where Assumption 4.(i) holds, we can use for any fixed $j \in \{2,\dots,n\}$ the splitting method to decompose the sum $\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)$ in different blocks whose lengths are given by the regeneration times of the split chain. Thanks to Assumption 4.(i), those blocks are independent and we can use standard concentration tools for sums of independent random variables. This approach is valid for any initial distribution of the chain. However, if Assumption 4.(i) is not satisfied, the blocks used to decompose $\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)$ are not independent and the splitting method can no longer be used. To bypass this issue, we need a Bernstein-type concentration inequality for additive functionals of Markov chains with time-dependent functions (see Proposition A.17 in Section 4.5.3). Proposition 4.19 is a straightforward corollary of [Jiang et al., 2018, Theorem 1] and requires Assumption 4.(ii) to be satisfied. We refer to Section 4.4.2 and in particular to Section 4.4.2.1 for further details.

4.3 Main results

4.3.1 Preliminary comments

Under the assumptions presented in Section 4.2, Theorem 4.3 and 4.4 provided in Section 4.3.2 give exponential inequalities for the U-statistic

$$U_{\mathrm{stat}}(n) = \sum_{1\leq i<j\leq n} \left(h_{i,j}(X_i, X_j) - \mathbb{E}\left[h_{i,j}(X_i, X_j)\right]\right).$$

Theorem 4.3 provides an Hoeffding-type concentration result that holds without any (or mild) condition on the initial distribution of the chain. By assuming that the chain $(X_i)_{i\geq 1}$ is stationary (meaning that X_1 is distributed according to π), Theorem 4.4 gives a Bernstein-type concentration inequality and leads to a better convergence rate compared to Theorem 4.3.
The proof of our main results relies on a martingale technique conducted by induction at depth $t_n := \lfloor r \log n \rfloor$ with $r > 2\left(\log(1/\rho)\right)^{-1}$ (see the remark following Assumption 1 for the definition of ρ). With the notations of Section 4.2, our concentration inequalities involve the following quantities

$$A := 2 \max_{i,j} \|h_{i,j}\|_\infty, \quad C_n^2 := \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathbb{E}\left[\mathbb{E}_{X'\sim\nu}[p_{i,j}^2(X_i, X')]\right], \tag{4.2}$$

$$B_n^2 := \max\Bigg[\max_{0\leq k\leq t_n} \max_i \sup_x \sum_{j=i+1}^{n} \mathbb{E}_{X'\sim\nu}\left(\mathbb{E}_{X\sim P^k(X',\cdot)}\, p_{i,j}(x, X)\right)^2, \max_{0\leq k\leq t_n} \max_j \sup_y \sum_{i=1}^{j-1} \mathbb{E}_{\tilde{X}\sim\pi}\left(\mathbb{E}_{X\sim P^k(y,\cdot)} p_{i,j}(\tilde{X}, X)\right)^2 \Bigg], \tag{4.3}$$

with the convention that $P^0(y,\cdot)$ is the Dirac measure at point $y \in E$. In the following, we will refer to those terms as *tail weights* for reasons that will become obvious after reading Section 4.3.2. Let us comment those terms.

- Understanding of the origin of B_n. B_n involves supremums over k ranging from 0 to t_n. The terms in the supremum corresponding to some specific value of k arise in our proof at the k-th step of our induction procedure (and will be denoted $\mathfrak{B}_k$ in Section 4.4, so that $B_n = \sup_{0\leq k\leq t_n} \mathfrak{B}_k$).
- Bounding B_n with uniform ergodicity. The uniform ergodicity of the Markov chain ensured by Assumption 1 can allow to bound the tail weight B_n since for all $x, y \in E$ and for all $k \geq 0$,

$$\left|\mathbb{E}_{X\sim P^k(y,\cdot)} p_{i,j}(x, X)\right| \leq \sup_z |h_{i,j}(x, z)| \times \|P^k(y,\cdot) - \pi\|_{\mathrm{TV}}.$$

- The case where $\nu = \pi$ and the independent setting
  In the specific case where $\nu = \pi$ (which includes the independent setting), we get that

$$C_n^2 = \sum_{i<j} \mathbb{E}\left\{\mathrm{Var}_{\tilde{X}\sim\pi}\left[h_{i,j}(X_i, \tilde{X})|X_i\right]\right\},$$

and using Jensen inequality that

$$B_n^2 \leq \max\Big[\sup_{x,i}\sum_{j=i+1}^{n}\mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(x,\tilde{X})]\,,\ \sup_{y,j}\sum_{i=1}^{j-1}\mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(\tilde{X},y)]\Big].$$

Hence, C_n^2 and B_n^2 can be understood as variance terms that would tend to be larger as ν moves away from π. Let us point out that in the independent setting, all terms for k ranging from 1 to t_n in the definition of B_n^2 vanish but the term corresponding to $k = 0$ does not since $P^0(y,\cdot)$ is the Dirac measure at y. We provide a detailed comparison of our results with known exponential inequalities in the independent setting in Section 4.3.3.2.

- Bounding B_n and C_n: A way to read immediately the convergence rates in our main results
  Using coarse bounds, one immediately gets that $B_n \leq A\sqrt{n}$ and $C_n \leq An$. We prompt the reader to keep in mind these bounds in order to directly see the rate of convergence and the dominant terms in the inequalities from Section 4.3.2. These bounds can be significantly improved for particular cases as done in the example presented in Section 4.3.3.5.

4.3.2 Exponential inequalities

We now state our first result Theorem 4.3 whose proof can be found in Section 4.4.1.1.

Theorem 4.3. *Let $n \geq 2$. We suppose Assumptions 1, 2 and 3 described in Section 4.2. There exist two constants $\beta, \kappa > 0$ such that for any $u > 0$,*

a) if Assumption 4.(i) is satisfied, it holds with probability at least $1 - \beta e^{-u}\log(n)$,

$$U_{\mathrm{stat}}(n) \leq \kappa\log(n)\Big(\left[A\log(n)\sqrt{n}\right]\sqrt{u} + \left[A + B_n\sqrt{n}\right]u + \left[2A\sqrt{n}\right]u^{3/2} + A\left[u^2+n\right]\Big),$$

b) if Assumption 4.(ii) is satisfied, it holds with probability at least $1 - \beta e^{-u}\log(n)$,

$$U_{\mathrm{stat}}(n) \leq \kappa\log(n)\Big(\left[\mathbf{C}_n + A\log(n)\sqrt{n}\right]\sqrt{u} + \left[A + B_n\sqrt{n}\right]u + \left[2A\sqrt{n}\right]u^{3/2} + A\left[u^2+n\right]\Big).$$

Note that the kernels $h_{i,j}$ do not need to be symmetric and that we do not consider any assumption on the initial measure of the Markov chain $(X_i)_{i\geq 1}$ if the kernels $h_{i,j}$ do not depend on i (see Assumption 4). By bounding coarsely B_n and C_n in Theorem 4.3 (respectively by $\sqrt{n}A$ and nA), we get that there exist constants $\beta, \kappa > 0$ such that for any $u \geq 1$, it holds with probability at least $1 - \beta e^{-u}\log n$,

$$\frac{2}{n(n-1)}U_{\mathrm{stat}}(n) \leq \kappa\max_{i,j}\|h_{i,j}\|_\infty \log n\left\{\frac{u}{n} + \left[\frac{u}{n}\right]^2\right\}. \tag{4.4}$$

In particular it holds

$$\frac{2}{n(n-1)}U_{\mathrm{stat}}(n) = \mathcal{O}_{\mathbb{P}}\left(\frac{\log(n)\log\log n}{n}\right),$$

where $\mathcal{O}_{\mathbb{P}}$ denotes stochastic boundedness. Up to a $\log(n)\log\log n$ multiplicative term, we uncover the optimal rate of Hoeffding's inequality for canonical U-statistics of order 2, see Joly and Lugosi [2016]. Taking a close look at the proof of Theorem 4.3 (and more specifically at Section 4.4.3), one can remark that the same results hold if the U-statistic is centered with the expectations $\mathbb{E}_\pi[h_{i,j}]$, namely for

$$\sum_{1\leq i<j\leq n}\left(h_{i,j}(X_i,X_j) - \mathbb{E}_\pi[h_{i,j}]\right).$$

It is well-known that one can expect a better convergence rate when variance terms are small with a Bernstein bound. The main limitation in Theorem 4.3 that prevents us from taking advantage of small variances is the term at the extreme right on the concentration inequality of Theorem 4.3, namely $An\log n$. Working with the additional assumption that the Markov chain $(X_i)_{i\geq 1}$ is stationary – meaning that the initial distribution of the chain is the stationary distribution π – we are able to prove a Bernstein-type concentration inequality as stated with Theorem 4.4. The proof of Theorem 4.4 is provided in Section 4.4.1.2. Stationarity is only used to bound the remaining terms that were not already considered in

the t_n steps of our induction procedure (see Section 4.3.1 for the definition of t_n). We refer to the proof of Proposition 4.6.b) in Section 4.4.3 for details.

Theorem 4.4. *We suppose Assumptions 1, 2 and 3 described in Section 4.2. We further assume that the Markov chain $(X_i)_{i\geq 1}$ is stationary. Then there exist two constants $\beta, \kappa > 0$ such that for any $u > 0$, it holds with probability at least $1 - \beta e^{-u} \log n$,*

$$U_{\text{stat}}(n) \leq \kappa \log(n) \Big(\left[C_n + A \log(n)\sqrt{n}\right] \sqrt{u} + \left[A + B_n\sqrt{n}\right] u + \left[2A\sqrt{n}\right] u^{3/2} + A\left[u^2 + \log \mathbf{n}\right] \Big).$$

In case where Assumption 4.(i) holds, one can remove C_n in the previous inequality.

4.3.3 Discussion

4.3.3.1 Examples of Markov chains satisfying the Assumptions

Example 1: Finite state space. For Markov chains with finite state space, Assumption 2 holds trivially. Hence, in such framework the result of Theorem 4.3 holds for any uniformly ergodic Markov chain. In particular, this is true for any aperiodic and irreducible Markov chains using [Behrends, 2000, Lemma 7.3.(ii)].

Example 2: AR(1) process. Let us consider the process $(X_n)_{n\in\mathbb{N}}$ on $\mathbb{R}^k$ defined by

$$X_0 \in \mathbb{R}^k \text{ and for all } n \in \mathbb{N}, \quad X_{n+1} = H(X_n) + Z_n,$$

where $(Z_n)_{n\in\mathbb{N}}$ are i.i.d random variables in $\mathbb{R}^k$ and $H : \mathbb{R}^k \to \mathbb{R}^k$ is an application. Such a process is called an auto-regressive process of order 1, noted AR(1). Assuming that the distribution of Z_1 has density f_Z with respect to the Lebesgue measure on $\mathbb{R}^k$ (denoted λ_{Leb}), it is well-known that mild regularity conditions on H and f_Z ensure that the Markov chain $(X_i)_{i\geq 1}$ is uniformly ergodic. These conditions require in particular that both f_Z and H are continuously differentiable with H bounded. We refer to Doukhan and Ghindès [1980] for the full statement.
We denote $B_H := B(0, \|H\|_\infty)$ the euclidean ball in $\mathbb{R}^k$ with radius $\|H\|_\infty$ centered at 0. Assuming that $y \mapsto \sup_{\{z\in B_H\}} f_Z(y - z)$ is integrable on $\mathbb{R}^k$ with respect to λ_{Leb}, we get that Assumption 2 holds (see the remark following Assumption 2). The previous condition on f_Z is for example satisfied for Gaussian distributions. We deduce that Theorems 4.3 and 4.4 can be applied in such settings that are typically found in nonlinear filtering problem (see [Del Moral and Guionnet, 1999, Section 4]).

Example 3: Arch process. Let us consider $E = \mathbb{R}$. The ARCH model is

$$X_{n+1} = H(X_n) + G(X_n)Z_{n+1},$$

where H and G are continuous functions, and $(Z_n)_n$ are i.i.d. centered normal random variables with variance $\sigma^2 > 0$. Assuming that $\inf_x |G(x)| \geq a > 0$, we know that the Markov chain $(X_n)_n$ is irreducible and aperiodic (see [Ango Nze, 1998, Lemma 1]). Assuming further that $\|H\|_\infty \leq b < \infty$ and that $\|G\|_\infty \leq c$, we can show that Assumptions 1 and 2 hold. Let us first remark that the transition kernel P of the Markov chain $(X_n)_n$ is such that for any $x \in \mathbb{R}$, $P(x, dy)$ has density $p(x, \cdot)$ with respect to the Lebesgue measure with

$$p(x, y) = \left(2\pi\sigma^2\right)^{-1} \exp\left(-\frac{(y - H(x))^2}{2\sigma^2 G(x)^2}\right).$$

Defining for any $y \in \mathbb{R}$,

$$g_m(y) := \frac{1}{2\pi\sigma^2} \times \begin{cases} \exp\left(-\frac{(y-B)^2}{2\sigma^2 a^2}\right) & \text{if } y < -b \\ \exp\left(-\frac{2B^2}{\sigma^2 a^2}\right) & \text{if } |y| \leq b \\ \exp\left(-\frac{(y+B)^2}{2\sigma^2 a^2}\right) & \text{if } y > b \end{cases}$$

$$\text{and} \quad g_M(y) := \left(2\pi\sigma^2\right)^{-1} \times \begin{cases} \exp\left(-\frac{(y+b)^2}{2\sigma^2 c^2}\right) & \text{if } y < -b \\ 1 & \text{if } |y| \leq b \\ \exp\left(-\frac{(y-b)^2}{2\sigma^2 c^2}\right) & \text{if } y > b \end{cases},$$

it holds $g_m(y) \leq p(x,y) \leq g_M(y)$ for any $x, y \in \mathbb{R}$. We deduce that considering $\delta_m = \|g_m\|_1, \delta_M = \|g_M\|_1$ and μ (resp. ν) with density $g_m/\|g_m\|_1$ (resp. $g_M/\|g_M\|_1$) with respect to the Lebesgue measure on $\mathbb{R}$, Assumptions 1 and 2 hold.

4.3.3.2 The independent setting

In the case where the random variables $(X_i)_{i\geq 1}$ are independent, the tail weights involved in our exponential inequality take the following form

$$C_n^2 = \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathbb{E}\left[p_{i,j}(X_i, X_j)^2\right] \quad \text{and} \quad B_n^2 = \max\left[\sup_{i,x} \sum_{j=i+1}^{n} \mathbb{E}\left[p_{i,j}^2(x, X_j)\right] \ , \ \sup_{j,y} \sum_{i=1}^{j-1} \mathbb{E}\left[p_{i,j}^2(X_i, y)\right]\right], \tag{4.5}$$

where we remind that all terms for $k \in \{1, \ldots, t_n\}$ in the definition of B_n^2 in Eq.(4.3) vanish and it only remains the contribution of terms for $k = 0$. In the independent setting, [Houdré and Reynaud-Bouret, 2003, Theorem 1] proved that for any $u > 0$, it holds with probability at least $1 - 3e^{-u}$,

$$U_{\text{stat}}(n) \leq C_n\sqrt{u} + (D_n + F_n)\, u + B_n u^{3/2} + Au^2,$$

where A, B_n and C_n coincide with the tail weights of our work (see Eq.(4.5)). Let us comment the tail weights involved in the different regimes of the tail behaviour.

- Sub-Gaussian. In Theorem 4.4, we recover the term C_n from Houdré and Reynaud-Bouret [2003] and we suffer an additional $A\sqrt{n}\log n$ term.
- Sub-Exponential. D_n and F_n come from duality arguments in the proof of Houdré and Reynaud-Bouret [2003]. We do not recover the counterpart of these terms in Theorem 4.4 since working with dependent variables bring additional technical difficulties and the use for example of a decoupling argument. $D_n + F_n$ is replaced by $A + B_n\sqrt{n}$ in our result.
- Sub-Weibull with parameter $2/3$. While Houdré and Reynaud-Bouret [2003] find the quantity B_n for the term $u^{3/2}$, the counterpart in Theorem 4.4 is the worst case scenario since it always holds $B_n \leq A\sqrt{n}$.
- Sub-Weibull with parameter $1/2$. We obtain the same behaviour for the sub-Weibull (with parameter $1/2$) regime of the tail behaviour.

Let us also mention that Theorem 4.4 has an additive term $A\log^2 n$ (that will not be dominant for standard choice of u). This term can be understood as a proof artefact and arises when we bound the remaining terms in the U-statistic that were not considered in our induction procedure. We finally point out that our result involves additive $\log n$ factors (both in the tail bound and in the probability).

4.3.3.3 Connections with the literature

In this section, we describe the concentration inequality obtained in Shen et al. [2020] for U-statistics in a dependent framework and we explain the differences with our work. We consider an integer $n \in \mathbb{N}\backslash\{0\}$ and a geometrically α-mixing sequence $(X_i)_{i\in[n]}$ (see [Shen et al., 2020, Section 2]) with coefficient

$$\alpha(i) \leq \gamma_1 \exp(-\gamma_2 i), \quad \text{for all } i \geq 1,$$

where γ_1, γ_2 are two positive absolute constants. We consider a kernel $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ degenerate, symmetric, continuous, integrable and satisfying for some $q \geq 1$, $\int_{\mathbb{R}^{2d}} |\mathcal{F}h(u)| \|u\|_2^q du < \infty$, where $\mathcal{F}h$ denotes the Fourier-transform of h. Then Eq.(2.4) from Shen et al. [2020] states that there exists a constant $c > 0$ such that for any $u > 0$, it holds with probability at least $1 - 6e^{-u}$

$$\frac{2}{n(n-1)} U_{\text{stat}}(n) \leq 4c \|\mathcal{F}h\|_{L^1} \left\{ A_n^{1/2} \frac{u}{n} + c \log^4(n) \left[\frac{u}{n}\right]^2 \right\}, \tag{4.6}$$

where $A_n^{1/2} = 4\left(\frac{64\gamma_1^{1/3}}{1-\exp(-\gamma_2/3)} + \frac{\log^4(n)}{n} \right)$ and $U_{\text{stat}}(n) = \sum_{1 \leq i < j \leq n} (h(X_i, X_j) - \mathbb{E}_\pi[h])$.
Shen et al. [2020] has the merit of working with geometrically α-mixing stationary sequences which includes in particular geometrically (and hence uniformly) ergodic Markov chains (see [Jones, 2004, p.6]). For the sake of simplicity, we presented the result of Shen et al. [2020] for U-statistics of order 2, but their result holds for U-statistics of arbitrary order $m \geq 2$. Nevertheless, they only consider state spaces like $\mathbb{R}^d$ with $d \geq 1$ and they work with a unique kernel h (i.e. $h_{i_1,\dots,i_m} = h$ for any $i_1, \dots, i_m$) which is assumed to be symmetric continuous, integrable and that satisfies some smoothness assumption. On the contrary, we consider general state spaces and we allow different kernels $h_{i,j}$ that are not assumed to be symmetric or smooth. In addition, Theorem 4.4 is a Bernstein-type exponential inequality where we can benefit from small variance terms, which is not the case for Shen et al. [2020]. We provide a specific example in Section 4.3.3.5. Table 4.1 summarizes the main differences beteen our results and the one from Shen et al. [2020].

	Shen et al. [2020]	**Our work**
State space	$\mathbb{R}^d$	General
Bernstein bound?	No	Yes
Hoeffding bound requires stationarity?	Yes	No
Order	Arbitrary m	$m = 2$
Dependence structure	Mixing condition	Uniform ergodicity
Assumptions on kernels	Unique, symmetric, smooth	Bounded

Table 4.1: Comparison between our concentration inequality and the existing literature.

4.3.3.4 Motivations for the study of time dependent kernels

In this section, we want to stress the importance of working with weighted U-statistics for practical applications. In the following, we detail two specific examples borrowed from the fields of information retrieval and of homogeneity tests. Note that one could find other applications such as in genetic association (cf. Wei et al. [2016]) or for independence tests (cf. Shieh et al. [1994]).

Average-Precision Correlation. When we search the Internet, the browser computes a numeric score on how well each object in the database matches the query, and rank the objects according to this value. In order to evaluate the quality of this browser, a standard approach in the field of information retrieval consists in comparing the ranking provided by the web search engine and the ranking obtained from human labels (cf. Han and Qian [2018]). One way to measure how well both rankings are aligned is to report the correlation between them. One of the most commonly used rank correlation statistic is the Kendall's τ. Considering a dataset of size $n \in \mathbb{N}$ ordered according to the human labels and denoting X_i the rank the browser gives to the i-th element, the Kendall's τ is defined by

$$\tau^{\text{Ken}} := \frac{2}{n(n-1)} \sum_{i \neq j} \left\{ \mathbb{1}_{X_i > X_j} \mathbb{1}_{i>j} + \mathbb{1}_{X_i < X_j} \mathbb{1}_{i<j} \right\} - 1.$$

Since only the top ranking objects are shown to the user, it would be legitimate to penalize heavier errors made on items having high rankings. The Kendall's τ does not make such distinctions and new correlation measurements have been popularized to address this issue. One of them is the so-called

Average-Precision Correlation (cf. Yilmaz et al. [2008]) which is defined by

$$\tau^{\mathrm{AP}} := \frac{2}{n-1}\sum_{j=2}^{n}\frac{\sum_{i=1}^{j-1}\mathbb{1}_{X_i<X_j}}{j-1} - 1.$$

Note that τ^{AP} is a U-statistic where the kernels $h_{i,j}(x,y) := \frac{\mathbb{1}_{x\le y}}{j-1}$ depend on j. Let us point out that $h_{i,j}$ do not depend on i so that Assumption 4.(i) holds.

Accounting for confounding covariates. U-statistics are powerful tools to compare the distributions of random variables across two groups (say with labels 0 and 1) from samples $X_1,\dots,X_n$ and $X_{n+1},\dots,X_{n+m}$. The typical example is the Wilcoxon Rank Sum Test (WRST) based on the following U-statistic

$$\sum_{i=1}^{n}\sum_{j=1}^{m} h(X_i, X_{n+j}) \quad \text{where} \quad h(x,y) := \frac{1}{2}\mathbb{1}_{x<y} + \frac{1}{2}\mathbb{1}_{x\le y}.$$

The WRST relies on the following idea: if the data is pooled and then ranked, the average rank of observations from each group should be the same. For any $i \in [n+m]$, let G_i be the random variable valued in $\{0,1\}$ allocating the i-th individual to one of the two groups. Note that the observed allocation $(g_i)_{i\in[n+m]}$ is given by $g_i = 0$ if and only if $i \le n$. When group membership is not assigned through randomization, there may be confounding covariates Z (assumed to be observed) that can cause a spurious association between outcome and group membership. In that case, we wish rather to test the null hypothesis $\mathbb{P}(X \le t \,|\, G=0, Z=z) = \mathbb{P}(X \le t \,|\, Z=z)$. In Satten et al. [2018], the authors developed such a test by working with the following adjusted U-statistics involving index-dependent kernels

$$\Big(\sum_{i\,:\,g_i=0} w(z_i,g_i)\sum_{j\,:\,g_j=1} w(z_j,g_j)\Big)^{-1}\sum_{i\,:\,g_i=0}\sum_{j\,:\,g_j=1} h(X_i,X_j)w(z_i,g_i)w(z_j,g_j),$$

where the weights $w(z_i,g_i) = (\mathbb{P}(G=g_i \,|\, Z=z_i))^{-1}$ can be estimated with a logistic regression.

4.3.3.5 Time dependent kernels and convergence rate

In this section, we consider a stationary Markov chain $(X_i)_{i\ge1}$ satisfying Assumptions 1, 2 and 3. We study the case where there exist reals $(a_{i,j})_{i,j\in\mathbb{N}}$ such that for all $i,j \in \mathbb{N}$, $h_{i,j} = a_{i,j}h$ for some π-canonical kernel $h : E^2 \to \mathbb{R}$. For simplicity, we consider that $\mathbb{E}_\pi h = 0$ leading to $p_{i,j} = h_{i,j}$. Let us consider the specific example where $a_{i,j} = |j-i|^{-1}$ for $i \ne j$. In such setting, the coefficients $a_{i,j}$'s are weighting each summand in the U-statistic: the larger $|j-i|$, the smaller is the contribution of the term indexed by (i,j) in the sum. As a result, interpreting indexes as time steps, the $a_{i,j}$'s can be understood as forgetting factors. Since

$$B_n^2 \le A^2 \max\Big\{\max_i \sum_{j=i+1}^{n}|j-i|^{-2}\,,\ \max_j\sum_{i=1}^{j-1}|j-i|^{-2}\Big\} \le A^2\sum_{j=2}^{n}|j-1|^{-2} \le A^2\frac{\pi^2}{6},$$

$$\text{and}\quad C_n^2 \le A^2\sum_{j=2}^{n}\sum_{i=1}^{j-1}|j-i|^{-2} \le A^2\sum_{s=1}^{n}\frac{s}{s^2} \le A^2\big(1+\int_1^n\frac{1}{x}dx\big) \le A^2(1+\log n),$$

Theorem 4.4 ensures that there exist constants $\beta,\kappa>0$ such that for any $u \ge 1$ it holds with probability at least $1-\beta e^{-u}\log n$,

$$\frac{2}{n(n-1)}U_{\mathrm{stat}}(n) \le \kappa A\log n\left(\log(n)\frac{\sqrt{u}}{n^{3/2}} + \Big[\frac{u}{n}\Big]^{3/2} + \Big[\frac{u}{n}\Big]^{2}\right).$$

In particular, with probability at least $1-\beta\frac{\log n}{n}$ we have $\frac{2}{n(n-1)}U_{\mathrm{stat}}(n) \le 3\kappa A\frac{\log^{5/2}n}{n^{3/2}}$. This convergence rate improves significantly the one obtained from an Hoeffding-type concentration inequality like Eq.(4.4) that would lead to $U_{\mathrm{stat}}(n) \le 2\kappa A\frac{\log^{3/2}n}{n}$ with probability at least $1-\beta\frac{\log n}{n}$.

4.4 Proofs

4.4.1 Proofs of Theorems 4.3 and 4.4

Our proof is inspired from Houdré and Reynaud-Bouret [2003] where a Bernstein-type inequality is shown for U-statistics of order 2 in the independent setting (note that the proof can also be found in [Giné and Nickl, 2016, Section 3.4.3]). Their proof relies on the *canonical* property of the kernel functions which endowed the U-statistic with a martingale structure. We want to use a similar argument and we decompose $U_{\text{stat}}(n)$ to recover the martingale property for each term (except for the last one). Considering for any $l \geq 1$ the σ-algebra $G_l = \sigma(X_1, \ldots, X_l)$, the notation $\mathbb{E}_l$ refers to the conditional expectation with respect to G_l. Then we decompose $U_{\text{stat}}(n)$ as follows,

$$U_{\text{stat}}(n) = M_{\text{stat}}^{(t_n)}(n) + R_{\text{stat}}^{(t_n)}(n), \tag{4.7}$$

$$\text{with} \quad M_{\text{stat}}^{(t_n)}(n) = \sum_{k=1}^{t_n} \sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)] \right),$$

$$R_{\text{stat}}^{(t_n)}(n) = \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}\left[h_{i,j}(X_i, X_j)\right] \right),$$

and where t_n is an integer that scales logarithmically with n. We recall that $t_n := \lfloor r \log n \rfloor$ with $r > 2\left(\log(1/\rho)\right)^{-1}$ where $\rho \in (0,1)$ is a constant characterizing the uniform ergodicity of the Markov chain (see Assumption 1). By convention, we assume here that for all $k < 1$, $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot]$. Hence the first term that we will consider is given by

$$U_n = \sum_{1 \leq i < j \leq n} h_{i,j}^{(0)}(X_i, X_{j-1}, X_j),$$

where for all $x, y, z \in E$, $h_{i,j}^{(0)}(x, y, z) = h_{i,j}(x, z) - \int_w h_{i,j}(x, w) P(y, dw)$.
We provide a detailed proof of a concentration result for U_n by taking advantage of its martingale structure. Reasoning by induction, we show that the $t_n - 1$ following terms involved in the decomposition (4.7) of $U_{\text{stat}}(n)$ can be handled using a similar approach. Since the last term $R_{\text{stat}}^{(t_n)}(n)$ of the decomposition (4.7) has not a martingale property, another argument is required. We deal with $R_{\text{stat}}^{(t_n)}(n)$ exploiting the uniform ergodicity of the Markov chain $(X_i)_{i \geq 1}$ which is guaranteed by Assumption 1 (see [Roberts and Rosenthal, 2004, Theorem 8]).
The cornerstones of our approach are the following two propositions whose proofs are postponed to Section 4.4.2 and Section 4.4.3 respectively.

Proposition 4.5. *Let $n \geq 2$. We keep the notations of Sections 4.2 and 4.3.1. We suppose Assumptions 1, 2 and 3 described in Section 4.2. There exist two constants $\beta, \kappa > 0$ such that for any $u > 0$,*

a) if Assumption 4.(i) is satisfied, it holds with probability at least $1 - \beta e^{-u}\log(n)$,

$$M_{\text{stat}}^{(t_n)}(n) \leq \kappa \log(n) \Big(\left[A\sqrt{n}\log n\right] \sqrt{u} + \left[A + B_n\sqrt{n}\right] u + \left[2A\sqrt{n}\right] u^{3/2} + Au^2 \Big).$$

b) if Assumption 4.(ii) is satisfied, it holds with probability at least $1 - \beta e^{-u}\log(n)$,

$$M_{\text{stat}}^{(t_n)}(n) \leq \kappa \log(n) \Big(\left[\mathbf{C}_n + A\sqrt{n}\log n\right] \sqrt{u} + \left[A + B_n\sqrt{n}\right] u + \left[2A\sqrt{n}\right] u^{3/2} + Au^2 \Big).$$

Proposition 4.6. *Let $n \geq 2$. We keep the notations of Sections 4.2 and 4.3.1. We suppose Assumptions 1, 2 and 3. Then*

a) $R_{\text{stat}}^{(t_n)}(n) \leq A\left(2L + nt_n\right)$.

b) if the Markov chain $(X_i)_{i \geq 1}$ is stationary, $R_{\text{stat}}^{(t_n)}(n) \leq 2LA\left(1 + t_n + t_n^2\right)$.

4.4.1.1 Proof of Theorem 4.3

We suppose Assumptions 1, 2, 3 and 4.(i) (respectively 4.(ii)). From the decomposition (4.7) coupled with Proposition 4.5.$a)$ (respectively Proposition 4.5.$b)$) and Proposition 4.6.$a)$, the result of Theo-

rem 4.3.a) (respectively Theorem 4.3.b)) is straightforward.

4.4.1.2 Proof of Theorem 4.4

We suppose Assumptions 1, 2 and 3. We assume in addition that the Markov chain is stationary which implies in particular that Assumption 4.(ii) holds. From the decomposition (4.7) coupled with Proposition 4.5.b) and Proposition 4.6.b), the result of Theorem 4.4 is straightforward. Note that in case Assumption 4.(i) holds, the quantity C_n (involved in the sub-Gaussian regime of the tail) can be removed from the inequality by simply using Proposition 4.5.a) rather than Proposition 4.5.b).

4.4.2 Proof of Proposition 4.5

Let us recall that Proposition 4.5 requires either a mild condition on the initial distribution of the Markov chain or the fact that the kernels $h_{i,j}$ do not depend on i (see Assumption 4). One only needs to consider different Bernstein concentration inequalities for sums of functions of Markov chains to go from one result to the other. In this section, we give the proof of Proposition 4.5 in the case where Assumption 4.(i) holds. We specify the part of the proof that should be changed to get the result when $h_{i,j}$ may depend on both i and j and when Assumption 4.(ii) holds. We make this easily identifiable using the symbol  .

4.4.2.1 Concentration of the first term of the decomposition of the U-statistic

Martingale structure of the U-statistic. Defining $Y_j = \sum_{i=1}^{j-1} h_{i,j}^{(0)}(X_i, X_{j-1}, X_j)$, U_n can be written as $U_n = \sum_{j=2}^{n} Y_j$. Since

$$\mathbb{E}_{j-1}[Y_j] = \mathbb{E}[Y_j \mid X_1, \ldots, X_{j-1}] = 0,$$

we know that $(U_k)_{k\geq 2}$ is a martingale relative to the σ-algebras G_l, $l \geq 2$. This martingale can be extended to $n = 0$ and $n = 1$ by taking $U_0 = U_1 = 0$, $G_0 = \{\emptyset, E\}$, $G_1 = \sigma(X_1)$. We will use the martingale structure of $(U_n)_n$ through the following Lemma.

Lemma 4.7. *(cf. [Giné and Nickl, 2016, Lemma 3.4.6])*
Let $(U_m, G_m), m \in \mathbb{N}$, be a martingale with respect to a filtration G_m such that $U_0 = U_1 = 0$. For each $m \geq 1$ and $k \geq 2$, define the angle brackets $A_m^k = A_m^k(U)$ of the martingale U by

$$A_m^k = \sum_{i=1}^{m} \mathbb{E}_{i-1}[(U_i - U_{i-1})^k]$$

(and note $A_1^k = 0$ for all k). Suppose that for $\alpha > 0$ and all $i \geq 1$, $\mathbb{E}[e^{\alpha|U_i - U_{i-1}|}] < \infty$. Then

$$\left(\epsilon_m := e^{\alpha U_m - \sum_{k\geq 2} \alpha^k A_m^k/k!}, G_m\right), \ m \in \mathbb{N},$$

is a supermartingale. In particular, $\mathbb{E}[\epsilon_m] \leq \mathbb{E}[\epsilon_1] = 1$, so that, if $A_m^k \leq w_m^k$ for constants $w_m^k \geq 0$; then

$$\mathbb{E}[e^{\alpha U_m}] \leq e^{\sum_{k\geq 2} \alpha^k w_m^k/k!}.$$

We will also use the following convexity result several times.

Lemma 4.8. *[Giné and Nickl, 2016, page 179] For all $\theta_1, \theta_2, \epsilon \geq 0$, and for all integer $k \geq 1$,*

$$(\theta_1 + \theta_2)^k \leq (1+\epsilon)^{k-1}\theta_1^k + (1+\epsilon^{-1})^{k-1}\theta_2^k.$$

For all $k \geq 2$ and $n \geq 1$, we have using Assumption 3:

$$\begin{aligned}
A_n^k &= \sum_{j=2}^{n} \mathbb{E}_{j-1}\left[\sum_{i=1}^{j-1} h_{i,j}^{(0)}(X_i, X_{j-1}, X_j)\right]^k \leq V_n^k := \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} h_{i,j}^{(0)}(X_i, X_{j-1}, X_j)\right|^k \\
&= \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(h_{i,j}(X_i, X_j) - \mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i, \tilde{X})] + \mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i, \tilde{X})] - \mathbb{E}_{j-1}[h_{i,j}(X_i, X_j)]\right)\right|^k \\
&= \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(p_{i,j}(X_i, X_j) + m_{i,j}(X_i, X_{j-1})\right)\right|^k,
\end{aligned}$$

$$\text{where} \quad p_{i,j}(x,z) = h_{i,j}(x,z) - \mathbb{E}_\pi[h_{i,j}] \quad \text{and} \quad m_{i,j}(x,y) = \mathbb{E}_\pi[h_{i,j}] - \int_z h_{i,j}(x,z)P(y,dz).$$

Using Lemma 4.8 with $\epsilon = 1/2$, we deduce that

$$\begin{aligned}
V_n^k &\leq \sum_{j=2}^{n} \mathbb{E}_{j-1}\left(\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right| + \left|\sum_{i=1}^{j-1} m_{i,j}(X_i, X_{j-1})\right|\right)^k \\
&\leq \left(\frac{3}{2}\right)^{k-1} \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^k + 3^{k-1} \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} m_{i,j}(X_i, X_{j-1})\right|^k.
\end{aligned}$$

Let us remark that

$$\begin{aligned}
\sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} m_{i,j}(X_i, X_{j-1})\right|^k &= \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(\mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i, \tilde{X})] - \mathbb{E}_{j-1}[h_{i,j}(X_i, X_j)]\right)\right|^k \\
&= \sum_{j=2}^{n}\left|\sum_{i=1}^{j-1}\left(\mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i, \tilde{X})] - \mathbb{E}_{j-1}[h_{i,j}(X_i, X_j)]\right)\right|^k \\
&= \sum_{j=2}^{n}\left|\mathbb{E}_{j-1}\left[\sum_{i=1}^{j-1}\left(\mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i, \tilde{X})] - h_{i,j}(X_i, X_j)\right)\right]\right|^k \\
&\leq \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(\mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i, \tilde{X})] - h_{i,j}(X_i, X_j)\right)\right|^k,
\end{aligned}$$

where the last inequality comes from Jensen's inequality. We obtain the following upper-bound for V_n^k,

$$V_n^k \leq 2 \times 3^{k-1} \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^k \leq 2 \times 3^{k-1}\delta_M \sum_{j=2}^{n} \mathbb{E}_{X_j'}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j')\right|^k,$$

where the random variables $(X_j')_j$ are i.i.d. with distribution ν (see Assumption 2). $\mathbb{E}_{X_j'}$ denotes the expectation on the random variable X_j'.

Lemma 4.9. *(cf. [Giné and Nickl, 2016, Ex.1 Section 3.4]) Let Z_j be independent random variables with respective probability laws P_j. Let $k > 1$, and consider functions $f_1, \ldots, f_N$ where for all $j \in [N]$, $f_j \in L^k(P_j)$. Then the duality of L^p spaces and the independence of the variables Z_j imply that*

$$\left(\sum_{j=1}^{N} \mathbb{E}\left[|f_j(Z_j)|^k\right]\right)^{1/k} = \sup_{\sum_{j=1}^{N} \mathbb{E}|\xi_j(Z_j)|^{k/(k-1)}=1} \sum_{j=1}^{N} \mathbb{E}\left[f_j(Z_j)\xi_j(Z_j)\right],$$

where the sup runs over $\xi_j \in L^{k/(k-1)}(P_j)$.

Then by the duality result of Lemma 4.9,

$$\begin{aligned}
\left(V_n^k\right)^{1/k} &\leq \left(2\delta_M \times 3^{k-1} \sum_{j=2}^{n} \mathbb{E}_{X_j'} \left| \sum_{i=1}^{j-1} p_{i,j}(X_i, X_j') \right|^k \right)^{1/k} \\
&\leq (2\delta_M)^{1/k} \sup_{\xi \in \mathcal{L}_k} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathbb{E}_{X_j'} \left[p_{i,j}(X_i, X_j') \xi_j(X_j') \right] \\
\text{where } \mathcal{L}_k &= \left\{ \xi = (\xi_2, \ldots, \xi_n) \; s.t. \; \forall 2 \leq j \leq n, \; \xi_j \in L^{k/(k-1)}(\nu) \; with \; \sum_{j=2}^{n} \mathbb{E}|\xi_j(X_j')|^{k/(k-1)} = 1 \right\}. \\
&= (2\delta_M)^{1/k} \sup_{\xi \in \mathcal{L}_k} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}_{X_j'} \left[p_{i,j}(X_i, X_j') \xi_j(X_j') \right]
\end{aligned}$$

Let us denote by F the subset of the set $\mathcal{F}(E, \mathbb{R})$ of all measurable functions from (E, Σ) to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that are bounded by A. We set $S := E \times F^{n-1}$. For all $i \in [n]$, we define W_i by

$$W_i := \big(X_i, \underbrace{0, \ldots, 0}_{(i-1) \text{ times}}, p_{i,i+1}(X_i, \cdot), p_{i,i+2}(X_i, \cdot), \ldots, p_{i,n}(X_i, \cdot)\big) \in S.$$

Hence for all $i \in [n]$, W_i is $\sigma(X_i)$-measurable. We define for any $\xi = (\xi_2, \ldots, \xi_n) \in \prod_{i=2}^{n} L^{k/(k-1)}(\nu)$ the function

$$\forall w = (x, p_2, \ldots, p_n) \in S, \quad f_\xi(w) = \sum_{j=2}^{n} \int p_j(y) \xi_j(y) d\nu(y).$$

Then setting $\mathcal{F} = \{f_\xi : \sum_{j=2}^{n} \mathbb{E}|\xi_j(X_j')|^{k/(k-1)} = 1\}$, we have

$$(V_n^k)^{1/k} \leq (2\delta_M)^{1/k} \sup_{f_\xi \in \mathcal{F}} \sum_{i=1}^{n-1} f_\xi(W_i).$$

By the separability of the L^p spaces of finite measures, $\mathcal{F}$ can be replaced by a countable subset $\mathcal{F}_0$. To upper-bound the tail probabilities of U_n, we will bound the variable V_n^k on sets of large probability using Talagrand's inequality. Then we will use Lemma 4.7 on these sets by means of optional stopping.

Application of Talagrand's inequality for Markov chains. The proof of Lemma 4.10 is provided in Section 4.5.1 and relies mainly of the Talagrand's inequality from [Samson, 2000, Theorem 3].

Lemma 4.10. *Let us denote*

$$Z = \sup_{f_\xi \in \mathcal{F}} \sum_{i=1}^{n-1} f_\xi(W_i), \quad \sigma_k^2 = \mathbb{E}\left[\sum_{i=1}^{n-1} \sup_{f_\xi \in \mathcal{F}} f_\xi(W_i)^2 \right] \quad \textit{and} \quad b_k = \sup_{w \in S} \sup_{f_\xi \in \mathcal{F}} |f_\xi(w)|.$$

Then it holds for any $t > 0$,

$$\mathbb{P}\left(Z > \mathbb{E}[Z] + t\right) \leq \exp\left(-\frac{1}{8\|\Gamma\|^2} \min\left(\frac{t^2}{4\sigma_k^2}, \frac{t}{b_k} \right) \right),$$

where Γ is a $n \times n$ matrix defined in Section 4.5.1 which satisfies $\|\Gamma\| \leq \frac{2L}{1-\rho}$.

Using Lemma 4.10, we deduce that for any $t > 0$,

$$\mathbb{P}\left((V_n^k)^{1/k} \geq (2\delta_M)^{1/k} \mathbb{E}[Z] + (2\delta_M)^{1/k} t \right) \leq \exp\left(-\frac{1}{8\|\Gamma\|^2} \min\left(\frac{t^2}{4\sigma_k^2}, \frac{t}{b_k} \right) \right),$$

which implies that for any $x \geq 0$,

$$\mathbb{P}\left((V_n^k)^{1/k} \geq (2\delta_M)^{1/k} \mathbb{E}[Z] + (2\delta_M)^{1/k} 2\sigma_k \sqrt{x} + (2\delta_M)^{1/k} b_k x \right) \leq \exp\left(-\frac{x}{8\|\Gamma\|^2} \right).$$

Using the change of variable $x = k8\|\Gamma\|^2 u$ with $u \geq 0$ in the previous inequality leads to

$$\mathbb{P}\left(\bigcup_{k=2}^{\infty}(V_n^k)^{1/k} \geq (2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k}\sigma_k 3\|\Gamma\|\sqrt{ku} + (2\delta_M)^{1/k}k8\|\Gamma\|^2 b_k u\right) \leq 1.62e^{-u},$$

because

$$1 \wedge \sum_{k=2}^{\infty}\exp\left(-ku\right) \leq 1 \wedge \frac{1}{e^u(e^u-1)} = \left(e^u \wedge \frac{1}{e^u-1}\right)e^{-u} \leq \frac{1+\sqrt{5}}{2}e^{-u} \leq 1.62e^{-u}.$$

Using Lemma 4.8 twice and using Holder inequality to bound b_k and σ_k^2, we obtain (4.9) from Lemma 4.11. The proof of Lemma 4.11 is postponed to Section 4.5.2.

Lemma 4.11. *For any $u > 0$, we denote*

$$\begin{aligned} w_n^k := &((1+\epsilon)^{k-1}2\delta_M\left(\mathbb{E}[Z]\right)^k + 2\delta_M(1+\epsilon^{-1})^{2k-2}\left(8\|\Gamma\|^2\right)^k(nA^2)A^{k-2}(ku)^k \\ &+ (1+\epsilon)^{k-1}(1+\epsilon^{-1})^{k-1}2\delta_M\left(3\|\Gamma\|\right)^k\mathfrak{B}_0^2A^{k-2}(nku)^{k/2}, \end{aligned}$$

$$\text{with } \mathfrak{B}_0^2 := \max\left[\max_i\Big\|\sum_{j=i+1}^{n}\mathbb{E}_{X\sim\nu}\left[p_{i,j}^2(\cdot,X)\right]\Big\|_\infty, \max_j\Big\|\sum_{i=1}^{j-1}\mathbb{E}_{X\sim\pi}[p_{i,j}^2(X,\cdot)]\Big\|_\infty\right] \leq B_n^2, \tag{4.8}$$

where the dependence in u of w_n^k is leaved implicit. Then it holds

$$\mathbb{P}\left(V_n^k \leq w_n^k \quad \forall k \geq 2\right) \geq 1 - 1.62e^{-u}. \tag{4.9}$$

Bounding $(\mathbb{E}[Z])^k$.

The way we bound $(\mathbb{E}[Z])^k$ is the only part of the proof that needs to be modified to get the concentration result when Assumption 4.(i) or Assumption 4.(ii) holds. This is where we can use different Bernstein concentration inequalities according to whether the splitting method is applicable or not (see Section 4.2.5 for details). Here we present the approach when $h_{i,j} \equiv h_{1,j},\ \forall i,j$ (i.e. when Assumption 4.(i) is satisfied). We refer to Section 4.5.3 for the details regarding the way we bound $(\mathbb{E}Z)^k$ when Assumption 4.(ii) holds.

Using Jensen inequality and Lemma 4.9, we obtain

$$\begin{aligned} (\mathbb{E}[Z])^k \leq \mathbb{E}[Z^k] &= \mathbb{E}\left[\left(\sup_{\xi\in\mathcal{L}_k}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}[p_{i,j}(X_i,X_j')\xi_j(X_j')]\right)^k\right] \\ &= \mathbb{E}\left[\sum_{j=2}^{n}\mathbb{E}_{X_j'}\left[\left|\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j')\right|^k\right]\right] = \sum_{j=2}^{n}\mathbb{E}\left[\left|\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j')\right|^k\right], \end{aligned}$$

where we recall that $\mathbb{E}_{X_j'}$ denotes the expectation on the random variable X_j'. One can remark that conditionally to X_j', the quantity $\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j')$ is a sum of function of the Markov chain $(X_i)_{i\geq 1}$. Hence to control this term, we apply a Bernstein inequality for Markov chains.
Let us consider some $j \in [n]$ and some $x \in E$. Using the notations of Section 4.2.3, we define

$$\forall l \in \{0,\dots,n\}, \quad Z_l^j(x) = \sum_{i=m(S_l+1)}^{m(S_{l+1}+1)-1}p_{i,j}(X_i,x).$$

By convention, we set $p_{i,j} \equiv 0$ for any $i \geq j$. Let us consider $N_j = \sup\{i \in \mathbb{N} \ : \ mS_{i+1} + m - 1 \leq j-1\}$.

Then using twice Lemma 4.8, we have

$$|\sum_{i=1}^{j-1} p_{i,j}(X_i,x)|^k = |\sum_{l=0}^{N_j} Z_l^j(x) + \sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,x)|^k \tag{4.10}$$

$$\leq \left(\frac{3}{2}\right)^{k-1} |\sum_{l=1}^{N_j} Z_l^j(x)|^k + 3^{k-1}|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,x)|^k$$

$$\leq \left(\frac{9}{4}\right)^{k-1} |\sum_{l=0}^{\lfloor N_j/2\rfloor} Z_{2l}^j(x)|^k + \left(\frac{9}{2}\right)^{k-1} |\sum_{l=0}^{\lfloor (N_j-1)/2\rfloor} Z_{2l+1}^j(x)|^k + 3^{k-1}|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,x)|^k. \tag{4.11}$$

We have $|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,x)| \leq AmT_{N_j+1}$. So using the definition of the Orlicz norm and the fact that the random variables $(T_i)_{i\geq 2}$ are i.i.d., it holds for any $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,x)\right| \geq t\right) \leq \mathbb{P}(T_{N_j+1} \geq \frac{t}{Am}) \leq \mathbb{P}(\max(T_1,T_2) \geq \frac{t}{Am})$$

$$\leq \mathbb{P}(T_1 \geq \frac{t}{Am}) + \mathbb{P}(T_2 \geq \frac{t}{Am}) \leq 4\exp(-\frac{t}{Am\tau}).$$

Hence, using that for an exponential random variable G with parameter 1, $\mathbb{E}[G^p] = p! \ \ \forall p \geq 0$,

$$\mathbb{E}\left[\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,x)\right|^k\right] = 4\int_0^{+\infty} \mathbb{P}\left(\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,x)\right|^k \geq t\right) dt$$

$$\leq 4\int_0^{+\infty} \exp(-\frac{t^{1/k}}{Am\tau}) \leq 4(Am\tau)^k \int_0^{+\infty} \exp(-v)kv^{k-1}dv = 4(Am\tau)^k k!,$$

The random variable $Z_{2l}^j(x)$ is $\sigma(X_{m(S_{2l}+1)},\ldots,X_{m(S_{2l+1}+1)-1})$-measurable. Let us insist that this holds because we consider that $h_{i,j} \equiv h_{1,j}, \ \forall i,j$ which implies that $p_{i,j} \equiv p_{1,j}, \ \forall i,j$. Hence for any $x \in E$, the random variables $(Z_{2l}^j(x))_l$ are independent (see Section 4.2.3). Moreover, one has that for any l, $\mathbb{E}[Z_{2l}^j(x)] = 0$. This is due to [Meyn and Tweedie, 1993, Eq.(17.23) Theorem 17.3.1] together with Assumption 3 which gives that $\forall x' \in E, \quad \mathbb{E}_{X\sim\pi}[p_{i,j}(X,x')] = 0$. Let us finally notice that for any $x \in E$ and any $l \geq 0$, $|Z_{2l}^j(x)| \leq AmT_{2l+1}$, so $\|Z_{2l}^j(x)\|_{\psi_1} \leq Am\max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1}) \leq Am\tau$. First, we use Lemma 4.12 to obtain that

$$\mathbb{E}\left|\sum_{l=0}^{\lfloor N_j/2\rfloor} Z_{2l}^j(x)\right|^k \leq \mathbb{E}\max_{0\leq s\leq n-1}\left|\sum_{l=0}^{s} Z_{2l}^j(x)\right|^k \leq 2\times 4^k \mathbb{E}\left|\sum_{l=0}^{n-1} Z_{2l}^j(x)\right|^k,$$

where for the last inequality we gathered (4.13) with the left hand side of (4.12) from Lemma 4.12.

Lemma 4.12. *(cf. [de la Pena and Giné, 2000, Lemma 1.2.6])*
Let us consider some separable Banach space B endowed with the norm $\|\cdot\|$. Let X_i, $i \leq n$, be independent centered B-valued random variables with norms L^p for some $p \geq 1$ and let ϵ_i be independent Rademacher random variables independent of the variables X_i. Then

$$2^{-p}\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i X_i\right\|^p \leq \mathbb{E}\left\|\sum_{i=1}^n X_i\right\|^p \leq 2^p\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i X_i\right\|^p, \tag{4.12}$$

$$\text{and} \quad \mathbb{E}\max_{k\leq n}\left\|\sum_{i=1}^k X_i\right\|^p \leq 2^{p+1}\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i X_i\right\|^p \tag{4.13}$$

Similarly, the random variables $(Z_{2l+1}^j(x))_l$ are independent and satisfy for any l, $\mathbb{E}[Z_{2l+1}^j(x)] = 0$. With

an analogous approach, we get that

$$\mathbb{E}\left|\sum_{l=0}^{\lfloor (N_j-1)/2\rfloor} Z_{2l+1}^j(x)\right|^k \leq \mathbb{E}\max_{0\leq s\leq n-1}\left|\sum_{l=0}^{s} Z_{2l+1}^j(x)\right|^k \leq 2\times 4^k\mathbb{E}\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(x)\right|^k.$$

Let us denote for any $j\in[n]$, $\mathbb{E}_{|X_j'}$ the conditional expectation with respect to the σ-algebra $\sigma(X_j')$. Coming back to (4.11), we proved that

$$\begin{aligned}
\mathbb{E}_{|X_j'}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i,X_j')\right|^k &\leq \left(\frac{9}{4}\right)^{k-1}\mathbb{E}_{|X_j'}\left|\sum_{l=0}^{\lfloor N_j/2\rfloor} Z_{2l}^j(X_j')\right|^k \\
&+\left(\frac{9}{2}\right)^{k-1}\mathbb{E}_{|X_j'}\left|\sum_{l=0}^{\lfloor (N_j-1)/2\rfloor} Z_{2l+1}^j(X_j')\right|^k + 3^{k-1}\mathbb{E}_{|X_j'}\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i,X_j')\right|^k \\
&\leq 2\times 9^k\mathbb{E}_{|X_j'}\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(X_j')\right|^k + 2\times 18^k\mathbb{E}_{|X_j'}\left|\sum_{l=0}^{n-1} Z_{2l}^j(X_j')\right|^k + 4(3Am\tau)^k k!. \qquad (4.14)
\end{aligned}$$

It remains to bound the two expectations in (4.14). The two latter expectations will be controlled similarly and we give the details for the first one. We use the following Bernstein's inequality with the sequence of random variables $(Z_{2l+1}^j(x))_l$.

Lemma 4.13. *(Bernstein's ψ_1 inequality, [Van Der Vaart and Wellner, 2013, Lemma 2.2.11] and the subsequent remark).*
If $Y_1,\dots,Y_n$ are independent random variables such that $\mathbb{E}Y_i=0$ and $\|Y_i\|_{\psi_1}\leq\tau$, then for every $t>0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n Y_i\right|>t\right)\leq 2\exp\left(-\frac{1}{K}\min\left(\frac{t^2}{n\tau^2},\frac{t}{\tau}\right)\right),$$

for some universal constant $K>0$ ($K=8$ fits).

We obtain

$$\mathbb{P}\left(\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(x)\right|>t\right)\leq 2\exp\left(-\frac{1}{K}\min\left(\frac{t^2}{nA^2m^2\tau^2},\frac{t}{Am\tau}\right)\right).$$

We deduce that for any $x\in E$, any $j\in[n]$ and any $t\geq 0$,

$$\mathbb{E}\left[\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(x)\right|^k\right]=\int_0^\infty \mathbb{P}\left(\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(x)\right|^k>t\right)dt = 2\int_0^\infty \exp\left(-\frac{1}{K}\min\left(\frac{t^{2/k}}{nA^2m^2\tau^2},\frac{t^{1/k}}{Am\tau}\right)\right)dt.$$

Let us remark that $\frac{t^{2/k}}{A^2m^2n\tau^2}\leq\frac{t^{1/k}}{Am\tau}\Leftrightarrow t\leq(nA\tau m)^k$. Hence for any $j\in[n]$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(X_j')\right|^k\right] \\
&\leq 2\int_0^{(nA\tau m)^k}\exp\left(-\frac{t^{2/k}}{KnA^2m^2\tau^2}\right)dt + 2\int_0^\infty\exp\left(-\frac{t^{1/k}}{KAm\tau}\right)dt. \\
&\leq 2\int_0^{n/K}\exp(-v)\frac{k}{2}v^{k/2-1}\left(\sqrt{K}n^{1/2}A\tau m\right)^k dv + 2\int_0^\infty\exp(-v)\,kv^{k-1}(KAm\tau)^k dv. \\
&\leq 2\int_0^{n/K}\exp(-v)\frac{k}{2}v^{k/2-1}\left(\sqrt{K}n^{1/2}A\tau m\right)^k dv + 2k\times(k-1)!(KAm\tau)^k \\
&\leq k\left(\sqrt{K}n^{1/2}A\tau m\right)^k\int_0^{n/K}\exp(-v)\,v^{k/2-1}dv + 2k!(KAm\tau)^k,
\end{aligned}$$

where we used again that if G is an exponential random variable with parameter 1, then for any $p\in$

$\mathbb{N}$, $\mathbb{E}[G^p] = p!$. Since for any real $l \geq 1$,

$$\int_0^{\frac{n}{K}} e^{-v} v^{l-1} dv = \sum_{r=0}^{+\infty} \frac{(-1)^r}{r!} \int_0^{\frac{n}{K}} v^{r+l-1} dv = \sum_{r=0}^{+\infty} \frac{(-1)^r}{r!} \frac{1}{r+l} \left(\frac{n}{K}\right)^{r+l}$$
$$\leq \left(\frac{n}{K}\right)^l \sum_{r=0}^{+\infty} \frac{(-1)^r}{r!} \frac{1}{l} \left(\frac{n}{K}\right)^r \leq \frac{(\frac{n}{K})^l}{l} e^{-\frac{n}{K}},$$

we get that

$$k \left(\sqrt{K} n^{1/2} A\tau m\right)^k \int_0^{\frac{n}{K}} \exp(-v)\, v^{k/2-1} dv \leq 2 \left(\sqrt{K} n^{1/2} A\tau m\right)^k e^{-n/K} \left(\frac{n}{K}\right)^{k/2} = 2 (nA\tau m)^k e^{-n/K}.$$

Hence we proved that for some universal constant $K > 1$,

$$\mathbb{E}\left[\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(x)\right|^k\right] \leq 2 (nA\tau m)^k e^{-n/K} + 2k!(KAm\tau)^k \leq 4k!(KAm\tau)^k,$$

since for all $k \geq 2$, $e^{-n/K}(n/K)^k/(k!) \leq 1$. Using a similar approach, one can show the same bound for the second expectation in (4.14). We proved that for some universal constant $K > 1$,

$$\begin{aligned}(\mathbb{E}[Z])^k &\leq \sum_{j=2}^{n} \mathbb{E}\left[\mathbb{E}_{|X'_j}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^k\right] \\ &\leq 2 \times 9^k \sum_{j=2}^{n} \mathbb{E}\left[\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(X'_j)\right|^k\right] + 2 \times 18^k \sum_{j=2}^{n} \mathbb{E}\left[\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{n-1} Z_{2l}^j(X'_j)\right|^k\right] + 4 \sum_{j=2}^{n} (3Am\tau)^k k! \\ &\leq 2n \times 18^k \times 4k!(KAm\tau)^k + 4n(3Am\tau)^k k! = 16n \times k!(KAm\tau)^k, \end{aligned} \tag{4.15}$$

where in the last inequality, we still call K the universal constant defined by $18K$.

Upper-bounding U_n using the martingale structure. Let

$$T + 1 := \inf\{l \in \mathbb{N} : V_l^k \geq w_n^k \text{ for some } k \geq 2\}.$$

Then, the event $\{T \leq l\}$ depends only on $X_1, \ldots, X_l$ for all $l \geq 1$. Hence, T is a stopping time for the filtration $(\mathcal{G}_l)_l$ where $\mathcal{G}_l = \sigma((X_i)_{i\in[l]})$ and we deduce that $U_l^T := U_{l\wedge T}$ for $l = 0, \ldots, n$ is a martingale with respect to $(\mathcal{G}_l)_l$ with $U_0^T = U_0 = 0$ and $U_1^T = U_1 = 0$. We remark that $U_j^T - U_{j-1}^T = U_j - U_{j-1}$ if $T \geq j$ and zero otherwise, and that $\{T \geq j\}$ is $\mathcal{G}_{j-1}$ measurable. Then, the angle brackets of this martingale admit the following bound:

$$\begin{aligned} A_n^k(U^T) &= \sum_{j=2}^{n} \mathbb{E}_{j-1}[(U_j^T - U_{j-1}^T)^k] \\ &\leq \sum_{j=2}^{n} \mathbb{E}_{j-1}|U_j - U_{j-1}|^k \mathbb{1}_{T\geq j} = \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} h^{(0)}(X_i, X_{j-1}, X_j)\right|^k \mathbb{1}_{T\geq j} \\ &= \sum_{j=2}^{n-1} V_j^k \mathbb{1}_{T=j} + V_n^k \mathbb{1}_{T\geq n} \leq w_n^k \left(\sum_{j=2}^{n-1} \mathbb{1}_{T=j} + \mathbb{1}_{T\geq n}\right) \leq w_n^k, \end{aligned}$$

since, by definition of T, $V_j^k \leq w_n^k$ for all k on $\{T \geq j\}$. Hence, Lemma 4.7 applied to the martingale U_n^T implies

$$\mathbb{E}e^{\alpha U_n^T} \leq \exp\left(\sum_{k\geq 2} \frac{\alpha^k}{k!} w_n^k\right).$$

Also, since V_n^k is nondecreasing in n for each k, inequality (4.9) implies that

$$\mathbb{P}(T < n) \leq \mathbb{P}\left(V_n^k \geq w_n^k \quad \text{for some } k \geq 2\right) \leq 1.62e^{-u}.$$

Thus we deduce that for all $s \geq 0$,

$$\mathbb{P}(U_n \geq s) \leq \mathbb{P}(U_n^T \geq s, T \geq n) + \mathbb{P}(T < n) \leq e^{-\alpha s} \exp\left(\sum_{k\geq 2} \frac{\alpha^k}{k!} w_n^k\right) + 1.62e^{-u}. \tag{4.16}$$

The final step of the proof consists in simplifying $\exp\left(\sum_{k\geq 2} \frac{\alpha^k}{k!} w_n^k\right)$.

$$\begin{aligned}\sum_{k\geq 2} \frac{\alpha^k}{k!} w_n^k = 2\delta_M \sum_{k\geq 2} \frac{\alpha^k}{k!}(1+\epsilon)^{k-1}(\mathbb{E}[Z])^k + 2\delta_M \sum_{k\geq 2} \frac{\alpha^k}{k!}(2+\epsilon+\epsilon^{-1})^{k-1}\left(3\|\Gamma\|\right)^k \mathfrak{B}_0^2 A^{k-2}(nku)^{k/2} \\ + 2\delta_M \sum_{k\geq 2} \frac{\alpha^k}{k!}(1+\epsilon^{-1})^{2k-2}\left(8\|\Gamma\|^2\right)^k (nA^2)A^{k-2}(ku)^k =: a_1 + a_2 + a_3. \end{aligned} \tag{4.17}$$

Using the bound Eq.(4.15) obtained on $(\mathbb{E}Z)^k$, Lemma 4.14 bounds the three sums a_1, a_2 and a_3.

Lemma 4.14. $\exp\left(\sum_{k\geq 2} \frac{\alpha^k}{k!} w_n^k\right) \leq \exp\left(\frac{\alpha^2 W^2}{1-\alpha c}\right)$ *where*

$$\begin{aligned} W &= 6\sqrt{\delta_M}(1+\epsilon)^{1/2} n^{1/2} KA\tau m \\ &\quad + \sqrt{2\delta_M}(2+\epsilon+\epsilon^{-1})^{1/2} 3\|\Gamma\|\mathfrak{B}_0\sqrt{nu} + \sqrt{2\delta_M}A(1+\epsilon^{-1})8\|\Gamma\|^2\sqrt{n}eu, \\ \textit{and} \quad c &= \max\left[(1+\epsilon)KA\tau m\ ,\ (2+\epsilon+\epsilon^{-1})\left(3\|\Gamma\|\right)A(nu)^{1/2}\ ,\ (1+\epsilon^{-1})^2\left(8\|\Gamma\|^2\right)Aeu\right]. \end{aligned}$$

Using the result from Lemma 4.14 in (4.16) and taking $s = 2W\sqrt{u} + cu$ and $\alpha = \sqrt{u}/(W + c\sqrt{u})$ in this inequality yields

$$\mathbb{P}\left(U_n \geq 2W\sqrt{u} + cu\right) \leq e^{-u} + 1.62e^{-u} \leq (1+e)e^{-u}.$$

By taking $\epsilon = 1/2$, we deduce that for any $u \geq 0$, it holds with probability at least $1 - (1+e)e^{-u}$

$$\begin{aligned}\sum_{i<j} h_j^{(0)}(X_i, X_{j-1}, X_j) \leq \quad & 12\sqrt{\delta_M}KA\tau m\sqrt{nu} + 18\sqrt{\delta_M}\|\Gamma\|\mathfrak{B}_0\sqrt{n}u + 100\sqrt{\delta_M}\|\Gamma\|^2 A\sqrt{n}eu^{3/2} \\ & + 3KA\tau mu + 27A\|\Gamma\|\sqrt{n}u^{3/2} + 72A\|\Gamma\|^2 eu^2,\end{aligned}$$

Denoting $\kappa := \max\left(12\sqrt{\delta_M}K\tau m\ ,\ 18\sqrt{\delta_M}\|\Gamma\|, 100\sqrt{\delta_M}\|\Gamma\|^2 e, 3K\tau m, 72\|\Gamma\|^2 e\right)$, we have with probability at least $1 - (1+e)e^{-u}$

$$\sum_{i<j} h_j^{(0)}(X_i, X_{j-1}, X_j) \leq \kappa\left(A\sqrt{n}\sqrt{u} + (A + \mathfrak{B}_0\sqrt{n})u + 2A\sqrt{n}u^{3/2} + Au^2\right).$$

4.4.2.2 Reasoning by descending induction with a logarithmic depth

As previously explained, we apply a proof similar to the one of the previous subsection on the $t_n := \lfloor r \log n \rfloor$ terms in the sum $M_{\text{stat}}^{(t_n)}(n)$ (see (4.7)), with $r > 2\left(\log(1/\rho)\right)^{-1}$. Let us give the key elements to justify such approach by considering the second term of the sum $M_{\text{stat}}^{(t_n)}(n)$, namely

$$\begin{aligned} &\sum_{i<j} \left(\mathbb{E}_{j-1}\left[h_{i,j}(X_i, X_j)\right] - \mathbb{E}_{j-2}\left[h_{i,j}(X_i, X_j)\right]\right) \\ = \ &\underbrace{\sum_{i=1}^{n-2}\sum_{j=i+2}^{n} h_{i,j}^{(1)}(X_i, X_{j-2}, X_{j-1})}_{=:U_{n-1}^{(1)}} + \underbrace{\sum_{i=1}^{n-1}\left\{\mathbb{E}_i\left[h_{i,i+1}(X_i, X_{i+1})\right] - \mathbb{E}_{i-1}\left[h_{i,i+1}(X_i, X_{i+1})\right]\right\}}_{=:(*)} \end{aligned} \tag{4.18}$$

where $h_{i,j}^{(1)}(x,y,z) = \int_w h_{i,j}(x,w)P(z,dw) - \int_w h_{i,j}(x,w)P^2(y,dw)$. Using McDiarmid's inequality for Markov chain (see [Paulin, 2015, Corollary 2.10 and Remark 2.11]), we obtain Lemma 4.15.

Lemma 4.15. *Let us consider $l \in \{1,\dots,t_n\}$. For any $u>0$, it holds with probability at least $1-2e^{-u}$,*

$$\left|\sum_{i=1}^{n-1}\sum_{j=i+1}^{(i+l)\wedge n}\left(\mathbb{E}_{j-l}\left[h_{i,j}(X_i,X_j)\right]-\mathbb{E}_{j-l-1}\left[h_{i,j}(X_i,X_j)\right]\right)\right| \leq 3At_n\sqrt{t_{mix}nu},$$

where t_{mix} is the mixing time of the Markov chain and is given by

$$t_{mix} := \min\left\{t\geq 0 \ : \ \sup_x \|P^t(x,\cdot)-\pi\|_{\mathrm{TV}} < \frac{1}{4}\right\}.$$

Lemma 4.15 allows to bound $(*)$ in (4.18) (by choosing $l=1$). Now we aim at proving a concentration result for the term

$$U_{n-1}^{(1)} = \sum_{j=2}^{n-1}\sum_{i=1}^{j-1} h_{i,j}^{(1)}(X_i,X_{j-1},X_j),$$

using an approach similar to the one of the previous subsection.

- Martingale structure
  Denoting $Y_j^{(1)} = \sum_{i=1}^{j-1} h_{i,j}^{(1)}(X_i,X_{j-1},X_j)$, we have $U_{n-1}^{(1)} = \sum_{j=2}^{n-1} Y_j^{(1)}$ which shows that $(U_n^{(1)})_n$ is a martingale with respect to the σ-algebras $(G_l)_l$. Indeed, we have $\mathbb{E}_{j-1}[Y_j^{(1)}]=0$.

- Talagrand's inequality To upper-bound $(V_n^k)_n$, we split it as previously namely

$$V_n^k := \sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} h_{i,j}^{(1)}(X_i,X_{j-1},X_j)\right|^k = \sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(I_{i,j}^{(1)}(X_i,X_j)-\mathbb{E}_{j-1}[I_{i,j}^{(1)}(X_i,X_j)]\right)\right|^k,$$

  where $I_{i,j}^{(1)}(x,z) = \int_w h_{i,j}(x,w)P(z,dw)$. Using as previously Lemma 4.8 with $\epsilon=1/2$, we get

$$\begin{aligned}
V_n^k &= \sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(I_{i,j}^{(1)}(X_i,X_j)-\mathbb{E}_{\tilde{X}\sim\pi}[I_{i,j}^{(1)}(X_i,\tilde{X})]\right.\right.\\
&\qquad\left.\left.+\mathbb{E}_{\tilde{X}\sim\pi}[I_{i,j}^{(1)}(X_i,\tilde{X})]-\mathbb{E}_{j-1}[I_{i,j}^{(1)}(X_i,X_j)]\right)\right|^k\\
&\leq (3/2)^{k-1}\sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(I_{i,j}^{(1)}(X_i,X_j)-\mathbb{E}_{\tilde{X}\sim\pi}[I_{i,j}^{(1)}(X_i,\tilde{X})]\right)\right|^k\\
&\quad + 3^{k-1}\sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(\mathbb{E}_{\tilde{X}\sim\pi}[I_{i,j}^{(1)}(X_i,\tilde{X})]-\mathbb{E}_{j-1}[I_{i,j}^{(1)}(X_i,X_j)]\right)\right|^k.
\end{aligned}$$

  Again, basic computations and Jensen's inequality lead to

$$\begin{aligned}
&\sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\left(\mathbb{E}_{\tilde{X}\sim\pi}[I_{i,j}^{(1)}(X_i,\tilde{X})]-\mathbb{E}_{j-1}[I_{i,j}^{(1)}(X_i,X_j)]\right)\right|^k\\
&=\sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}^{(1)}(X_i,X_j)\right|^k,
\end{aligned}$$

  where $p_{i,j}^{(1)}(x,z) := I_{i,j}^{(1)}(x,z) - \mathbb{E}_{\tilde{X}\sim\pi}[I_{i,j}^{(1)}(x,\tilde{X})]$. Hence, using Assumption 1 and Lemma 4.12

exactly like in the previous section, we get (for $(X'_j)_j$ i.i.d. with distribution ν)

$$V_n^k = 2 \times 3^{k-1} \sum_{j=2}^{n-1} \mathbb{E}_{j-1} \left| \sum_{i=1}^{j-1} p_{i,j}^{(1)}(X_i, X_j) \right|^k \leq 2 \times 3^{k-1} \delta_M \sum_{j=2}^{n-1} \mathbb{E}_{X'_j} \left| \sum_{i=1}^{j-1} p_{i,j}^{(1)}(X_i, X'_j) \right|^k .$$

Then, one can use the same duality trick to show that the V_n^k can be controlled using the supremum of a sum of functions of the Markov chain $(X_i)_{i\geq 1}$ using [Samson, 2000, Theorem 3].

- Bounding $\exp(w_n^k \alpha^k / k!)$
  The terms a_2 and a_3 (see Eq.(4.17)) can be bounded in a similar way. For the term a_1, we only need to show that $p_{i,j}^{(1)}$ satisfies $\mathbb{E}_{X_i \sim \pi} | p_{i,j}^{(1)}(X_i, z)] = 0, \ \forall z \in E$ in order to apply as previously a Bernstein's type inequality.

$$\begin{aligned} \mathbb{E}_{X_i \sim \pi} | p_{i,j}^{(1)}(X_i, z)] &= \int_{x_i} d\pi(x_i) \int_w h_{i,j}(x_i, w) P(z, dw) - \mathbb{E}_{X \sim \pi} \mathbb{E}_{\tilde{X} \sim \pi} [I_{i,j}^{(1)}(X, \tilde{X})] \\ &= \mathbb{E}_\pi[h_{i,j}] - \mathbb{E}_\pi[h_{i,j}] \quad \text{(Using Assumption 3)} \\ &= 0. \end{aligned}$$

- Conclusion of the proof
  Let us consider the quantities A_1 and $\mathfrak{B}_1$ defined as the counterparts of A and $\mathfrak{B}_0$ (see (4.8)) by replacing the functions $(p_{i,j})_{i,j}$ by $\left(p_{i,j}^{(1)}\right)_{i,j}$. One can easily see that $A_1 = A$. Let us give details about $\mathfrak{B}_1$.

  For any $x \in E$,

$$\begin{aligned} \mathbb{E}_{X' \sim \nu} \left[(p_{i,j}^{(1)})^2(x, X') \right] &= \int_z \left(I_{i,j}^{(1)}(x, z) - \mathbb{E}_{\tilde{X} \sim \pi} [I_{i,j}^{(1)}(x, \tilde{X})] \right)^2 d\nu(z) \\ &= \int_z \left(\int_w h_{i,j}(x, w) P(z, dw) - \int_w h_{i,j}(x, w) \underbrace{\int_a P(a, dw) d\pi(a)}_{= d\pi(w)} \right)^2 d\nu(z) \\ &= \mathbb{E}_{X' \sim \nu} \left[\mathbb{E}_{X \sim P(X', \cdot)} h_{i,j}(x, X) - \mathbb{E}_\pi[h_{i,j}] \right]^2 , \end{aligned}$$

  and for any $y \in E$,

$$\begin{aligned} \mathbb{E}_{\tilde{X} \sim \pi} \left[(p_{i,j}^{(1)})^2(\tilde{X}, y) \right] &= \int_x \left(I_{i,j}^{(1)}(x, y) - \mathbb{E}_{\tilde{X} \sim \pi} [I_{i,j}^{(1)}(x, \tilde{X})] \right)^2 d\pi(x) \\ &= \int_x \left(\int_w h_{i,j}(x, w) P(y, dw) - \int_w h_{i,j}(x, w) \underbrace{\int_a P(a, dw) d\pi(a)}_{= d\pi(w)} \right)^2 d\pi(x) \\ &= \mathbb{E}_{\tilde{X} \sim \pi} \left[\mathbb{E}_{X \sim P(y, \cdot)} h_{i,j}(\tilde{X}, X) - \mathbb{E}_\pi[h_{i,j}] \right]^2 . \end{aligned}$$

  Hence we get that

$$\mathfrak{B}_1^2 := \max \left[\max_i \left\| \sum_{j=i+1}^{n} \mathbb{E}_{X \sim \nu} \left[(p_{i,j}^{(1)})^2(\cdot, X) \right] \right\|_\infty , \max_j \left\| \sum_{i=1}^{j-1} \mathbb{E}_{X \sim \pi} [(p_{i,j}^{(1)})^2(X, \cdot)] \right\|_\infty \right] \leq B_n^2, \quad (4.19)$$

where we recall that

$$B_n^2 = \max\left[\sup_{0\le k\le t_n}\max_i \sup_x \sum_{j=i+1}^{n} \mathbb{E}_{X'\sim\nu}\left[\mathbb{E}_{X\sim P^k(X',\cdot)}h_{i,j}(x,X) - \mathbb{E}_\pi[h_{i,j}]\right]^2, \right.$$
$$\left. \sup_{0\le k\le t_n}\max_j \sup_y \sum_{i=1}^{j-1} \mathbb{E}_{\tilde{X}\sim\pi}\left[\mathbb{E}_{X\sim P^k(y,\cdot)}h_{i,j}(\tilde{X},X) - \mathbb{E}_\pi[h_{i,j}]\right]^2\right].$$

This allows us to get a concentration inequality similar to the one of the previous subsection, namely for any $u>0$, it holds with probability at least $1-(1+e)e^{-u}$,

$$\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} h_{i,j}^{(1)}(X_i,X_{j-1},X_j) \le \kappa\left(A\sqrt{n}\sqrt{u} + (A+B_n\sqrt{n})u + 2A\sqrt{n}u^{3/2} + Au^2\right)$$

Going back to (4.18) and using Lemma 4.15, we get that for any $u>0$, it holds with probability at least $1-(1+e+2)e^{-u}$,

$$\begin{aligned}&\sum_{i<j}\left(\mathbb{E}_{j-1}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}_{j-2}\left[h_{i,j}(X_i,X_j)\right]\right)\\ \le\ & \kappa\left(A\sqrt{n}\sqrt{u} + (A+B_n\sqrt{n})u + 2A\sqrt{n}u^{3/2} + Au^2\right) + 3At_n\sqrt{t_{mix}nu}\end{aligned} \tag{4.20}$$

One can do the same analysis for the t_n first terms in the decomposition (4.7). Still denoting κ the constant $\kappa + 3\sqrt{t_{mix}}$, we get that for any $u>0$ it holds with probability at least $1-(3+e)e^{-u}t_n$,

$$M_{\text{stat}}^{(t_n)}(n) \le \kappa t_n\left(At_n\sqrt{n}\sqrt{u} + (A+B_n\sqrt{n})u + 2A\sqrt{n}u^{3/2} + Au^2\right).$$

4.4.3 Proof of Proposition 4.6

In the following, we assume that $t_n \le n$, otherwise $R_{\text{stat}}^{(t_n)}(n)$ is an empty sum. Using our convention which states that for all $k<1$, $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot]$, we need to control

$$\left|R_{\text{stat}}^{(t_n)}(n)\right| = \left|\sum_{i<j}\left(\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right]\right)\right| \le (1)+(2), \tag{4.21}$$

with denoting $H_{i,j} = \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right]$,

$$(1) := \left|\sum_{i=1}^{n-t_n}\sum_{j=i+t_n}^{n} H_{i,j}\right| = \left|\sum_{j=t_n+1}^{n}\sum_{i=1}^{j-t_n} H_{i,j}\right|$$
$$\text{and } (2) := \left|\sum_{i=1}^{n-1}\sum_{j=i+1}^{(i+t_n-1)\wedge n} H_{i,j}\right| = \left|\sum_{j=2}^{n}\sum_{i=(j-t_n+1)\vee 1}^{j-1} H_{i,j}\right|.$$

We start by bounding the term (1) regardless of the initial distribution of the chain. We will bound in different ways the term (2) depending on whether the Markov chain is stationary or not. Let us first bound the term (1) splitting it into two terms,

$$(1) = \left|\sum_{j=t_n+1}^{n}\sum_{i=1}^{j-t_n}\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right]\right| \le (1a)+(1b).$$

Using Assumption 3, it holds $\mathbb{E}_\pi[h_{i,j}] = \mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i,\tilde{X})] = \int_x h_{i,j}(X_i,x)d\pi(x)$. Hence we get that

$$\begin{aligned}
(1a) &:= \left| \sum_{j=t_n+1}^{n} \sum_{i=1}^{j-t_n} \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}_\pi[h_{i,j}] \right| \\
&\leq \sum_{j=t_n+1}^{n} \left| \int_{x_j} \sum_{i=1}^{j-t_n} h_{i,j}(X_i,x_j)\left(P^{t_n}(X_{j-t_n},dx_j) - d\pi(x_j)\right) \right| \\
&\leq \sum_{j=t_n+1}^{n} \sup_{x_j} \left| \sum_{i=1}^{j-t_n} h_{i,j}(X_i,x_j) \right| \sup_z \|P^{t_n}(z,\cdot) - \pi\|_{\mathrm{TV}} \\
&\leq \sum_{j=t_n+1}^{n} \sup_{x_j} \left| \sum_{i=1}^{j-t_n} h_{i,j}(X_i,x_j) \right| L\rho^{t_n} \leq \sum_{j=t_n+1}^{n} \sup_{x_j} \left| \sum_{i=1}^{j-t_n} h_{i,j}(X_i,x_j) \right| L\frac{1}{n^2} \leq LA,
\end{aligned}$$

where in the penultimate inequality we used that $\rho^{t_n} \leq \rho^{r\log(n)} = n^{r\log(\rho)} \leq n^{-2}$. Indeed $2+r\log(\rho) < 0$ because we choose r such that $r > 2(\log(1/\rho))^{-1}$.
Using again Assumption 3, it holds $\mathbb{E}_\pi[h_{i,j}] = \int_{x_i} \chi P^i(dx_i) \int_x h_{i,j}(x_i,x)d\pi(x)$ where χ is the initial distribution of the Markov chain $(X_i)_{i\geq 1}$. We get that

$$\begin{aligned}
(1b) &:= \left| \sum_{j=t_n+1}^{n} \sum_{i=1}^{j-t_n} \mathbb{E}_\pi[h_{i,j}] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right] \right| \\
&\leq \sum_{j=t_n+1}^{n} \sum_{i=1}^{j-t_n} \left| \int_{x_i}\int_{x_j} h_{i,j}(x_i,x_j)\chi P^i(dx_i)\left(P^{j-i}(x_i,dx_j) - d\pi(x_j)\right) \right| \\
&\leq \sum_{j=t_n+1}^{n} \sum_{i=1}^{j-t_n} \|h_{i,j}\|_\infty \underbrace{\int_{x_i} \chi P^i(dx_i)}_{=1} \sup_z \int_{x_j} \left| P^{j-i}(z,dx_j) - d\pi(x_j) \right| \\
&\leq \sum_{j=t_n+1}^{n} \sum_{i=1}^{j-t_n} \|h_{i,j}\|_\infty L\rho^{j-i} \leq \sum_{j=t_n+1}^{n} \sum_{i=1}^{j-t_n} \|h_{i,j}\|_\infty L\rho^{t_n} \leq LA,
\end{aligned}$$

where in the penultimate inequality we used that $\rho^{t_n} \leq \rho^{r\log(n)} = n^{r\log(\rho)} \leq n^{-2}$.

4.4.3.1 Bounding (2) without stationarity

Without assuming that the Markov chain is stationary, we bound coarsely (2) as follows

$$(2) = \left| \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-1} \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right] \right| \leq Ant_n.$$

This concludes the proof of Proposition 4.6.a) since we obtain, $R_{\mathrm{stat}}^{(t_n)}(n) \leq A\left(2L + nt_n\right)$.

4.4.3.2 Bounding (2) with stationarity

Considering now that the chain is stationary, we split (2) into three different contributions.

$$(2) = \left| \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-1} \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right] \right| \leq (2a)+(2b)+(2c),$$

$$\text{with} \quad (2a) := \left| \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}_{\pi}\left[h_{i,j}\right] \right|,$$

$$(2b) := \left| \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \mathbb{E}_{\pi}\left[h_{i,j}\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right] \right|,$$

$$\text{and} \quad (2c) := \left| \sum_{j=2}^{n} \sum_{i=(j-\lfloor \frac{t_n}{2} \rfloor+1)\vee 1}^{j-1} \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right] \right|.$$

The only place where we use the stationarity of the chain is to bound the terms $(2b)$ and $(2c)$ by writing that $\mathbb{E}[h_{i,j}(X_i,X_j)] = \int_{x_i}\int_{x_j} d\pi(x_i)P^{j-i}(x_i,dx_j)$. Both are bounded using similar ideas, that is why we show here how to deal with $(2b)$ and we postpone the proof of Lemma 4.16 to Section 4.5.5.

Lemma 4.16. *It holds* $\quad (2a) \leq LAt_n \quad$ *and* $\quad (2c) \leq 2LAt_n^2$.

We show now how we deal with the term $(2b)$.

$$\begin{aligned}
(2b) &:= \left| \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \mathbb{E}_{\pi}\left[h_{i,j}\right] - \mathbb{E}\left[h_{i,j}(X_i,X_j)\right] \right| \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \left| \int_{x_i}\int_{x_j} h_{i,j}(x_i,x_j) d\pi(x_i)\left(d\pi(x_j) - P^{j-i}(x_i,dx_j)\right) \right| \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \|h_{i,j}\|_\infty \underbrace{\int_{x_i} d\pi(x_i)}_{=1} \underbrace{\sup_y \int_{x_j} \left|d\pi(x_j) - P^{j-i}(y,dx_j)\right|}_{=\sup_y \|P^{j-i}(y,\cdot)-\pi\|_{\mathrm{TV}}} \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \|h_{i,j}\|_\infty L\rho^{j-i} \leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \|h_{i,j}\|_\infty L\rho^{t_n/2} \leq LAt_n,
\end{aligned}$$

where we used that $\rho^{t_n/2} \leq \rho^{r\log(n)/2} = n^{r\log(\rho)/2} \leq n^{-1}$. Indeed $1+r\log(\rho)/2 < 0$ because we choose r such that $r > 2(\log(1/\rho))^{-1}$. Coming back to Eq.(4.21), we deduce that $R_{\mathrm{stat}}^{(t_n)}(n) \leq AL\left(2+2t_n+2t_n^2\right)$ which concludes the proof of Proposition 4.6.

4.5 Proofs of technical Lemmas

4.5.1 Proof of Lemma 4.10

In the section, we show that in the proof of Proposition 4.5, we can use the concentration inequality for the supremum of an empirical process of [Samson, 2000, Theorem 3].
Let us consider the sequence of random variables $W = (W_1,\dots,W_n)$ on the probability space $(\Omega,\mathcal{A},\mathbb{P})$ taking values in the measurable space $S = E \times F^{n-1}$ where F is the subset of the set $\mathcal{F}(E,\mathbb{R})$ of all measurable functions from (E,Σ) to $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ that are bounded by A. Note that

$$\{0_{\mathcal{F}(E,\mathbb{R})}\} \cup \{p_{i,j}(x,\cdot) \ : \ x \in E, \ i,j \in [n]\} \subset F.$$

We define $\mathcal{P} := \left\{D \in \mathcal{P}(F) \ : \ \forall i \in [n-1], \forall j \in \{i+1,\dots,n\}, \ f_{i,j}^{-1}(D) \in \Sigma\right\}$ where $\mathcal{P}(F)$ is the power-

set of F and where $\forall i \in [n-1],\ \forall j \in \{i+1,\dots,n\}$,

$$f_{i,j} : (E,\Sigma) \to (F,\mathcal{P}(F)) \qquad x \mapsto p_{i,j}(x,\cdot).$$

Then we have the following straightforward result.

Lemma 4.17. *$\mathcal{P}$ is a σ-algebra on F.*

In the following, we endow the space F with the σ-algebra $\mathcal{P}$ and we consider on S the product σ-algebra given by $\mathcal{S} := \sigma\left(\{C \times D_2 \times \cdots \times D_n \ :\ C \in \Sigma,\ D_j \in \mathcal{P}\ \forall j \in \{2,\dots,n\}\}\right)$.
For all $i \in [n]$, we define W_i by

$$W_i := \big(X_i,\ \underbrace{0,\dots,0}_{(i-1)\text{ times}}, p_{i,i+1}(X_i,\cdot), p_{i,i+2}(X_i,\cdot),\dots,p_{i,n}(X_i,\cdot)\big).$$

Hence for all $i \in [n]$, W_i is $\sigma(X_i)$-measurable. Let us consider for any $i \in [n-1]$,

$$\Phi_i : (E,\Sigma) \to (S,\mathcal{S}) \qquad \text{such that} \quad \forall x \in E,\ \Phi_i(x) = \big(x, \underbrace{0_F,\dots,0_F}_{(i-1)\text{ times}}, p_{i,i+1}(x,\cdot),\dots,p_{i,n}(x,\cdot)\big).$$

Then, one can directly see that for all $i \in [n-1]$, $W_i = \Phi_i(X_i)$ and by construction of $\mathcal{P}$ and $\mathcal{S}$, Φ_i is measurable. Indeed, each coordinate of Φ_i is measurable by construction of $\mathcal{P}$ and this ensures that Φ_i is measurable thanks to the following Lemma.

Lemma 4.18. *(cf. [Aliprantis and Border, 2006, Lemma 4.49]) Let (X,Σ), (X_1,Σ_1) and (X_2,Σ_2) be measurable spaces, and let $f_1 : X \to X_1$ and $f_2 : X \to X_2$. Define $f : X \to X_1 \times X_2$ by $f(x) = (f_1(x), f_2(x))$. Then $f : (X,\Sigma) \to (X_1 \times X_2, \Sigma_1 \otimes \Sigma_2)$ is measurable if and only if the two functions $f_1 : (X,\Sigma) \to (X_1,\Sigma_1)$ and $f_2 : (X,\Sigma) \to (X_2,\Sigma_2)$ are both measurable.*

Then it holds for any $i \in \{2,\dots,n-1\}$ and any $G \in \mathcal{S}$,

$$\begin{aligned}&\mathbb{P}\left(W_i \in G \mid W_{i-1}\right) = \mathbb{P}\left(\Phi_i(X_i) \in G \mid W_{i-1}\right) = \mathbb{P}\left(\Phi_i(X_i) \in G \mid X_{i-1}\right)\\ &= \mathbb{P}\left(X_i \in \Phi_i^{-1}(G) \mid X_{i-1}\right) = P\left(X_{i-1}, \Phi_i^{-1}(G)\right) = \left[(\Phi_i)_{\#}P(X_{i-1},\cdot)\right](G), \end{aligned} \tag{4.22}$$

where $(\Phi_i)_{\#}P(X_{i-1},\cdot)$ denotes the pushforward measure of the measure $P(X_{i-1},\cdot)$ by the measurable map Φ_i. We deduce that W_i is non-homogeneous Markov chain. Moreover, (4.22) proves that the transition kernel of the Markov chain $(W_k)_k$ from state $i-1$ to state i is given by $K^{(i-1,i)}$ where for all $(x,p_2,\dots,p_n) \in S$ and for all $G \in \mathcal{S}$,

$$K^{(i-1,i)}((x,p_2,\dots,p_n),G) = \left[(\Phi_i)_{\#}P(x,\cdot)\right](G).$$

One can easily generalize this notation. Let us consider some $i,j \in [n]$ with $i < j$ and let us denote by $K^{(i,j)}$ the transition kernel of the Markov chain $(W_k)_k$ from state i to state j. Then for all $x \in E$, for all $p_2,\dots,p_n \in F$ and for all $G \in \mathcal{S}$,

$$K^{(i,j)}((x,p_2,\dots,p_n),G) = \left[(\Phi_j)_{\#}P^{j-i}(x,\cdot)\right](G),$$

We introduce the mixing matrix $\Gamma = (\gamma_{i,j})_{1\le i,j\le n-1}$ where coefficients are defined by

$$\gamma_{i,j} := \sup_{w_i \in S}\sup_{z_i \in S} \|\mathcal{L}(W_j|W_i = w_i) - \mathcal{L}(W_j|W_i = z_i)\|_{\mathrm{TV}}.$$

For any $w \in S = E \times F^{n-1}$, we denote by $w^{(1)}$ the first coordinate of the vector w. Hence, $w^{(1)}$ is an

element of E. Then

$$\begin{aligned}
\gamma_{i,j} &= \sup_{w_i\in S}\sup_{z_i\in S}\sup_{G\in\mathcal{S}}\left|\left[(\Phi_j)_{\#}P^{j-i}(w_i^{(1)},\cdot)\right](G)-\left[(\Phi_j)_{\#}P^{j-i}(z_i^{(1)},\cdot)\right](G)\right| \\
&= \sup_{w_i\in S}\sup_{z_i\in S}\sup_{G\in\mathcal{S}}\left|P^{j-i}\left(w_i^{(1)},\Phi_j^{-1}(G)\right)-P^{j-i}\left(z_i^{(1)},\Phi_j^{-1}(G)\right)\right| \\
&\leq \sup_{w_i\in S}\sup_{z_i\in S}\sup_{C\in\Sigma}\left|P^{j-i}\left(w_i^{(1)},C\right)-P^{j-i}\left(z_i^{(1)},C\right)\right| \\
&= \sup_{x_i\in E}\sup_{x_i'\in E}\sup_{C\in\Sigma}\left|P^{j-i}(x_i,C)-P^{j-i}(x_i',C)\right| \\
&= \sup_{x_i\in E}\sup_{x_i'\in E}\|P^{j-i}(x_i,\cdot)-\pi(\cdot)+\pi(\cdot)-P^{j-i}(x_i',\cdot)\|_{\mathrm{TV}} \\
&\leq \sup_{x_i\in E}\|P^{j-i}(x_i,\cdot)-\pi(\cdot)\|_{\mathrm{TV}}+\sup_{x_i'\in E}\|P^{j-i}(x_i',\cdot)-\pi(\cdot)\|_{\mathrm{TV}}\leq 2L\rho^{j-i},
\end{aligned}$$

where in the first inequality we used that $\Phi_j:(E,\Sigma)\to(S,\mathcal{S})$ is measurable and in the last inequality we used the uniform ergodicity of the Markov chain $(X_i)_{i\geq 1}$. We deduce that

$$\|\Gamma\|\leq 2L\left\|\mathrm{Id}+\sum_{l=1}^{n-1}\rho^l N_l\right\|,\quad N_l=\left(n_{i,j}^{(l)}\right)_{1\leq i,j\leq n-1}\text{ with }n_{i,j}^{(l)}=\begin{cases}1 & \text{if } j-i=l\\ 0 & \text{otherwise.}\end{cases}.$$

Note that N_l is a nilpotent matrix of order l. Since for each $1\leq l\leq n-1$, $\|N_l\|\leq 1$, it follows from the triangular inequality that

$$\|\Gamma\|\leq 2L\sum_{l=0}^{n-1}\rho^l\leq\frac{2L}{1-\rho}.$$

To conclude the proof and get the concentration result stated in Lemma 4.10, one only needs to apply [Samson, 2000, Theorem 3] with the class of functions $\mathcal{F}$ and with the Markov chain $(W_k)_k$. Let us recall that $\mathcal{F}$ is defined by $\mathcal{F}=\{f_\xi:\sum_{j=2}^n\mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1\}$ where for any $\xi=(\xi_2,\dots,\xi_n)\in\prod_{i=2}^n L^{k/(k-1)}(\nu)$,

$$\forall w=(x,p_2,\dots,p_n)\in E\times F^{n-1},\quad f_\xi(w)=\sum_{j=2}^n\int p_j(y)\xi_j(y)d\nu(y).$$

4.5.2 Proof of Lemma 4.11

Bounding b_k. Using Hölder's inequality we have,

$$\begin{aligned}
b_k &= \sup_{w\in S}\sup_{f_\xi\in\mathcal{F}}|f_\xi(w)| = \sup_{(p_2,\dots,p_n)\in F^{n-1}}\sup_{\xi\in\mathcal{L}_k}\sum_{j=2}^n\mathbb{E}[p_j(X_j')\xi_j(X_j')] \\
&\leq \sup_{(p_2,\dots,p_n)\in F^{n-1}}\sup_{\sum_{j=2}^n\mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1}\sum_{j=2}^n\left(\mathbb{E}\left|p_j(X_j')\right|^k\right)^{1/k}\left(\mathbb{E}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
&\leq \sup_{(p_2,\dots,p_n)\in F^{n-1}}\sup_{\sum_{j=2}^n\mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1}\left(\sum_{j=2}^n\mathbb{E}\left|p_j(X_j')\right|^k\right)^{1/k}\left(\sum_{j=2}^n\mathbb{E}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
&\leq \sup_{(p_2,\dots,p_n)\in F^{n-1}}\left(\sum_{j=2}^n\mathbb{E}\left|p_j(X_j')\right|^k\right)^{1/k}\leq((nA^2)A^{k-2})^{1/k},
\end{aligned}$$

where $A:=2\max_{i,j}\|h_{i,j}\|_\infty$ which satisfies $\max_{i,j}\|p_{i,j}\|_\infty\leq A$. Here, we used that F is the set of measurable functions from (E,Σ) to $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ bounded by A.

Bounding the variance.

$$\sigma_k^2 = \mathbb{E}\left[\sum_{i=1}^{n-1} \sup_{f_\xi \in \mathcal{F}} f_\xi(W_i)^2\right] = \sum_{i=1}^{n-1} \mathbb{E}\left[\sup_{\xi\in\mathcal{L}_k}\left(\sum_{j=i+1}^{n} \mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]\right)^2\right]$$
$$= \sum_{i=1}^{n-1} \mathbb{E}\left[\left(\sup_{\xi\in\mathcal{L}_k}\left|\sum_{j=i+1}^{n} \mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]\right|\right)^2\right] \le n\left(\mathfrak{B}_0^2 A^{k-2}\right)^{2/k},$$

where the last inequality comes from the following (where we use twice Holder's inequality),

$$\begin{aligned}
&\sup_{\xi\in\mathcal{L}_k}\left|\sum_{j=i+1}^{n} \mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]\right| \\
\le\ & \sup_{\xi\in\mathcal{L}_k}\sum_{j=i+1}^{n}\left(\mathbb{E}_{X_j'}\left|p_{i,j}(X_i,X_j')\right|^k\right)^{1/k}\left(\mathbb{E}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
\le\ & \sup_{\sum_{j=2}^n \mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1}\left(\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}\left|p_{i,j}(X_i,X_j')\right|^k\right)^{1/k}\left(\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
\le\ & \left(\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}\left|p_{i,j}(X_i,X_j')\right|^k\right)^{1/k} \le \left(\mathfrak{B}_0^2 A^{k-2}\right)^{1/k}, \quad \text{where } \mathfrak{B}_0 \text{ is defined in (4.8).}
\end{aligned}$$

Using Lemma 4.8 twice and the bounds obtained on b_k and σ_k^2 gives for $u>0$,

$$\begin{aligned}
&\left[(2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k}\sigma_k 3\|\Gamma\|\sqrt{ku} + (2\delta_M)^{1/k}k8\|\Gamma\|^2 b_k u\right]^k \\
\le\ & \left[(2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k}3\|\Gamma\|(\mathfrak{B}_0^2A^{k-2})^{1/k}\sqrt{nku} + (2\delta_M)^{1/k}8\|\Gamma\|^2((nA^2)A^{k-2})^{1/k}ku\right]^k \\
\le\ & (1+\epsilon)^{k-1}2\delta_M\left(\mathbb{E}[Z]\right)^k + (1+\epsilon^{-1})^{k-1}\Big[(2\delta_M)^{1/k}8\|\Gamma\|^2((nA^2)A^{k-2})^{1/k}ku \\
& + (2\delta_M)^{1/k}3\|\Gamma\|(\mathfrak{B}_0^2A^{k-2})^{1/k}\sqrt{nku}\Big]^k \\
\le\ & (1+\epsilon)^{k-1}2\delta_M\left(\mathbb{E}[Z]\right)^k + 2\delta_M(1+\epsilon^{-1})^{2k-2}\left(8\|\Gamma\|^2\right)^k(nA^2)A^{k-2}(ku)^k \\
& + (1+\epsilon)^{k-1}(1+\epsilon^{-1})^{k-1}2\delta_M\left(3\|\Gamma\|\right)^k\mathfrak{B}_0^2A^{k-2}(nku)^{k/2}.
\end{aligned}$$

4.5.3 Bounding $(\mathbb{E}[Z])^k$ under Assumption 4.(ii)

In this section, we only provide the part of the proof of Proposition 4.5 that needs to be modified to get the result when the kernels $h_{i,j}$ depend on both i and j and when Assumption 4.(ii) holds. Keeping the notations of the proof of Proposition 4.5, we only want to bound $(\mathbb{E}[Z])^k$ (and thus a_1) using a different concentration result that can allow to deal with kernel functions $h_{i,j}$ that might depend on i. We will use Proposition 4.19 which is proved in the Appendix A.5.2.

Proposition 4.19. *(cf. Proposition A.17) Suppose that the sequence $(X_i)_{i\ge1}$ is a Markov chain satisfying Assumptions 1 and 4.(ii) with stationary distribution π and with an absolute spectral gap $1-\lambda>0$. Let us consider some $n\in\mathbb{N}^*$ and bounded real valued functions $(f_i)_{1\le i\le n}$ such that for any $i\in\{1,\dots,n\}$, $\int f_i(x)d\pi(x)=0$ and $\|f_i\|_\infty\le c$ for some $c>0$. Let $\sigma^2=\sum_{i=1}^n\int f_i^2(x)d\pi(x)/n$. Then for any $0\le t<(1-\lambda)/(5cq)$,*

$$\mathbb{E}_\chi\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] \le \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\exp\left(\frac{n\sigma^2}{qc^2}(e^{tqc}-tqc-1)+\frac{n\sigma^2\lambda qt^2}{1-\lambda-5cqt}\right), \tag{4.23}$$

where q is the constant introduced in Assumption 4.(ii). Moreover for any $u \geq 0$ it holds

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f_i(X_i) > \frac{2quA_1c}{n} + \sqrt{\frac{2quA_2\sigma^2}{n}}\right) \leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} e^{-u}.$$

where $A_2 := \frac{1+\lambda}{1-\lambda}$ and $A_1 := \frac{1}{3}\mathbb{1}_{\lambda=0} + \frac{5}{1-\lambda}\mathbb{1}_{\lambda>0}$.

Note that Proposition 4.19 is an extension of the Bernstein inequality from [Jiang et al., 2018, Theorem 1] where the authors only consider stationary chains. Following the approach used in [Fan et al., 2021, Theorem 2.3], we show in Section A.5.2 that we can extend their result to obtain Proposition 4.19 by working under the milder assumption Assumption 4.(ii). Let us recall that

$$\begin{aligned}
(\mathbb{E}[Z])^k \leq \mathbb{E}[Z^k] \quad &\text{(Using Jensen's inequality)} \\
&= \mathbb{E}\left[\left(\sup_{f_\xi\in\mathcal{F}}\sum_{i=1}^{n-1} f_\xi(X_i)\right)^k\right] = \mathbb{E}\left[\left(\sup_{f_\xi\in\mathcal{F}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}_{j-1}[p_{i,j}(X_i,X_j')\xi_j(X_j')]\right)^k\right] \\
&= \mathbb{E}\left[\sum_{j=2}^{n}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i,X_j')\right|^k\right] \quad \text{(using Lemma 4.9)} \\
&= \mathbb{E}\left[\sum_{j=2}^{n}\mathbb{E}_{|X'}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i,X_j')\right|^k\right].
\end{aligned}$$

Thus we have

$$a_1 = \frac{2\delta_M}{1+\epsilon}\mathbb{E}\sum_{j=2}^{n}\left(\mathbb{E}_{|X'}\left[e^{\alpha(1+\epsilon)K|C^{(j)}|}\right] - \alpha(1+\epsilon)K\mathbb{E}_{|X'}\left[|C^{(j)}|\right] - 1\right),$$

where $C^{(j)} = \sum_{i=1}^{j-1} p_{i,j}(X_i,X_j')$ and where the notation $\mathbb{E}_{|X'}$ refers to the expectation conditionally to the σ-algebra $\sigma(X_2',\dots,X_n')$.
Now we use a symmetrization trick: since $e^x - x - 1 \geq 0$ for all x and since $e^{a|x|} + e^{-a|x|} = e^{ax} + e^{-ax}$, adding $\mathbb{E}_{|X'}[\exp\left(-\alpha(1+\epsilon)K|C^{(j)}|\right)] + \alpha(1+\epsilon)K\mathbb{E}_{|X'}[|C^{(j)}|] - 1$ to a_1 gives

$$a_1 \leq \frac{2\delta_M}{1+\epsilon}\mathbb{E}\sum_{j=2}^{n}\left(\mathbb{E}_{|X'}[e^{\alpha(1+\epsilon)KC^{(j)}}] - 1 + \mathbb{E}_{|X'}[e^{-\alpha(1+\epsilon)KC^{(j)}}] - 1\right). \tag{4.24}$$

Let us consider some $j \in \{2,\dots,n\}$. Conditionally on $\sigma(X_2',\dots,X_n')$, $C^{(j)}$ is a sum of bounded functions (by A) depending on the Markov chain. We denote

$$v_j(X_j') = \sum_{i=1}^{j-1}\mathbb{E}_{X_i\sim\pi}[p_{i,j}^2(X_i,X_j')|X_j'] \leq \mathfrak{B}_0^2$$

and $V = \sum_{j=2}^{n}\mathbb{E}v_j^k(X_j') \leq C_n^2\mathfrak{B}_0^{2(k-1)}$ (with $C_n^2 = \sum_{j=2}^{n}\sum_{i=1}^{j-1}\mathbb{E}[p_{i,j}^2(X_i,X_j')]$).
Remark that

$$\begin{aligned}
\mathbb{E}_{X_i\sim\pi}[p_{i,j}(X_i,X_j')|X_j'] &= \mathbb{E}_{X_i\sim\pi}\left[h_{i,j}(X_i,X_j') - \mathbb{E}_{\tilde{X}\sim\pi}[h_{i,j}(X_i,\tilde{X})] \mid X_j'\right] \\
&= \int_{x'}\left(\int_{x_i}(h_{i,j}(x_i,X_j') - h_{i,j}(x_i,\tilde{x}))d\pi(x_i)\right)d\pi(\tilde{x}) = 0,
\end{aligned}$$

where the last equality comes from Assumption 3. We use the Bernstein inequality for Markov chain (see Proposition 4.19). Notice from Taylor expansion that $(1-p/3)(e^p - p - 1) \leq p^2/2$ for all $p \geq 0$. Applying (4.23) with $t = \alpha(1+\epsilon)K$ and $c = A$, we get that for $\alpha < [(1+\epsilon)K\sqrt{q}(A\sqrt{q}/3 + \mathfrak{B}_0\sqrt{3/2})]^{-1} \wedge$

$[(1-\lambda)^{-1/2}(1+\epsilon)K\sqrt{q}\left(5A\sqrt{q}(1-\lambda)^{-1/2}+\sqrt{3\lambda}\mathfrak{B}_0\right)]^{-1}$,

$$\mathbb{E}_{|X'}[e^{\alpha(1+\epsilon)K|C^{(j)}|}] \leq 2\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} \times \mathbb{E}_{|X'}\left[\exp\left(\frac{\alpha^2(1+\epsilon)^2K^2qv_j(X_j')}{2-2Aq\alpha(1+\epsilon)K/3}+\frac{v_j(X_j')\lambda\alpha^2(1+\epsilon)^2K^2q}{1-\lambda-5\alpha(1+\epsilon)KAq}\right)\right].$$

Considering $\alpha < [(1+\epsilon)K\sqrt{q}(A\sqrt{q}/3+\mathfrak{B}_0\sqrt{3/2})]^{-1} \wedge [(1-\lambda)^{-1/2}(1+\epsilon)K\sqrt{q}\left(5A\sqrt{q}(1-\lambda)^{-1/2}+\sqrt{3\lambda}\mathfrak{B}_0\right)]^{-1}$, $\epsilon < 1$ and using (4.24), this leads to

$$\begin{aligned}
\frac{a_1}{2\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}} &\leq \frac{2\delta_M}{1+\epsilon}\sum_{j=2}^{n}\mathbb{E}\left[\exp\left(\frac{\alpha^2(1+\epsilon)^2K^2qv_j(X_j')}{2-2Aq\alpha(1+\epsilon)K/3}+\frac{v_j(X_j')\lambda\alpha^2(1+\epsilon)^2K^2q}{1-\lambda-5\alpha(1+\epsilon)KAq}\right)-1\right]\\
&= \frac{2\delta_M}{1+\epsilon}\sum_{j=2}^{n}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{\alpha^2(1+\epsilon)^2K^2qv_j(X_j')}{2-2Aq\alpha(1+\epsilon)K/3}+\frac{v_j(X_j')\lambda\alpha^2(1+\epsilon)^2K^2q}{1-\lambda-5\alpha(1+\epsilon)KAq}\right)^k\\
&= \frac{2\delta_M}{1+\epsilon}\sum_{j=2}^{n}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{3}{2}\right)^{k-1}\left(\frac{\alpha^2(1+\epsilon)^2K^2qv_j(X_j')}{2-2Aq\alpha(1+\epsilon)K/3}\right)^k\\
&\quad+\frac{2\delta_M}{1+\epsilon}\sum_{j=2}^{n}\sum_{k=1}^{\infty}\frac{1}{k!}3^{k-1}\left(\frac{v_j(X_j')\lambda\alpha^2(1+\epsilon)^2K^2q}{1-\lambda-5\alpha(1+\epsilon)KAq}\right)^k \quad \text{(using Lemma 4.8)}\\
&\leq \frac{\delta_M}{3(1+\epsilon)}\sum_{k=1}^{\infty}\frac{3^k\alpha^{2k}(1+\epsilon)^{2k}K^{2k}q^kV}{(4-4Aq\alpha(1+\epsilon)K/3)^k}+\frac{2\delta_M}{3(1+\epsilon)}\sum_{k=1}^{\infty}\frac{3^kV\lambda^k\alpha^{2k}(1+\epsilon)^{2k}K^{2k}q^k}{(1-\lambda-5\alpha(1+\epsilon)KAq)^k}\\
&\leq \frac{\delta_M}{3(1+\epsilon)}\sum_{k=1}^{\infty}\frac{3^k\alpha^{2k}(1+\epsilon)^{2k}K^{2k}q^kC_n^2\mathfrak{B}_0^{2(k-1)}}{(2-2Aq\alpha(1+\epsilon)K/3)^k}+\frac{2\delta_M}{3(1+\epsilon)}\sum_{k=1}^{\infty}\frac{3^kC_n^2\mathfrak{B}_0^{2(k-1)}\lambda^k\alpha^{2k}(1+\epsilon)^{2k}K^{2k}q^k}{(1-\lambda-5\alpha(1+\epsilon)KAq)^k}\\
&= \frac{(1+\epsilon)C_n^2\alpha^2K^2\delta_Mq}{2-2Aq\alpha(1+\epsilon)K/3-3\alpha^2(1+\epsilon)^2K^2\mathfrak{B}_0^2q}+\frac{2\delta_MC_n^2\lambda\alpha^2(1+\epsilon)K^2q}{1-\lambda-5\alpha(1+\epsilon)KAq-3\mathfrak{B}_0^2\lambda\alpha^2(1+\epsilon)^2K^2q}\\
&= \frac{(1+\epsilon)C_n^2\alpha^2K^2\delta_Mq/2}{1-Aq\alpha(1+\epsilon)K/3-3\alpha^2(1+\epsilon)^2K^2\mathfrak{B}_0^2q/2}\\
&\qquad+\frac{2\delta_MC_n^2\lambda\alpha^2(1+\epsilon)K^2q(1-\lambda)^{-1}}{1-5(1-\lambda)^{-1}\alpha(1+\epsilon)KAq-3\mathfrak{B}_0^2\lambda(1-\lambda)^{-1}\alpha^2(1+\epsilon)^2K^2q}\\
&\leq \frac{(1+\epsilon)C_n^2\alpha^2K^2\delta_Mq/2}{1-\alpha(1+\epsilon)K\sqrt{q}(A\sqrt{q}/3+\mathfrak{B}_0\sqrt{3/2})}\\
&\qquad+\frac{2\delta_MC_n^2\lambda\alpha^2(1+\epsilon)K^2q(1-\lambda)^{-1}}{1-\alpha(1-\lambda)^{-1/2}(1+\epsilon)K\sqrt{q}\left(5A\sqrt{q}(1-\lambda)^{-1/2}+\sqrt{3\lambda}\mathfrak{B}_0\right)}.
\end{aligned}$$

From this bound on a_1, one can follow the steps of the proof of Proposition 4.5 to conclude.

4.5.4 Proof of Lemma 4.14

Bounding a_3. Using the inequality $k! \geq (k/e)^k$, we have,

$$\begin{aligned}
a_3 &\leq 2\delta_M\sum_{k\geq 2}\alpha^k(1+\epsilon^{-1})^{2k-2}\left(8\|\Gamma\|^2\right)^k(nA^2)A^{k-2}(eu)^k\\
&= 2\delta_M\alpha^2\left[\sqrt{n}A(1+\epsilon^{-1})8\|\Gamma\|^2eu\right]^2\sum_{k\geq 2}\alpha^{k-2}(1+\epsilon^{-1})^{2(k-2)}\left(8\|\Gamma\|^2\right)^{k-2}A^{k-2}(eu)^{k-2}\\
&= \frac{2\delta_M\alpha^2\left[\sqrt{n}A(1+\epsilon^{-1})8\|\Gamma\|^2eu\right]^2}{1-\alpha(1+\epsilon^{-1})^2\left(8\|\Gamma\|^2\right)Aeu}, \quad \text{for } \alpha < \left((1+\epsilon^{-1})^2\left(8\|\Gamma\|^2\right)Aeu\right)^{-1}.
\end{aligned}$$

Bounding a_2. We use the inequality $k! \geq k^{k/2}$ because $(k/e)^k > k^{k/2}$ for $k \geq e^2$ and for k smaller, the

inequality follows by direct verification. Hence,

$$
\begin{aligned}
a_2 &\leq 2\delta_M \sum_{k\geq 2} \alpha^k (2+\epsilon+\epsilon^{-1})^{k-1} \left(3\|\Gamma\|\right)^k \mathfrak{B}_0^2 A^{k-2} (nu)^{k/2} \\
&= 2\delta_M (2+\epsilon+\epsilon^{-1}) \alpha^2 \left[3\|\Gamma\|\mathfrak{B}_0\sqrt{nu}\right]^2 \sum_{k\geq 2} \alpha^{k-2} (2+\epsilon+\epsilon^{-1})^{k-2} \left(3\|\Gamma\|\right)^{k-2} A^{k-2} (nu)^{(k-2)/2} \\
&= \frac{2\delta_M (2+\epsilon+\epsilon^{-1}) \alpha^2 \left[3\|\Gamma\|\mathfrak{B}_0\sqrt{nu}\right]^2}{1-\alpha(2+\epsilon+\epsilon^{-1})\left(3\|\Gamma\|\right) A (nu)^{1/2}}, \quad \text{for } \alpha < \left((2+\epsilon+\epsilon^{-1})\left(3\|\Gamma\|\right) A (nu)^{1/2}\right)^{-1}.
\end{aligned}
$$

Bounding a_1**.** Using the bound previously obtained for $(\mathbb{E}[Z])^k$ we get,

$$
\begin{aligned}
a_1 &= 2\delta_M \sum_{k\geq 2} \frac{\alpha^k}{k!} (1+\epsilon)^{k-1} (\mathbb{E}[Z])^k \leq 32\delta_M n \sum_{k\geq 2} \alpha^k (1+\epsilon)^{k-1} (KAm\tau)^k \\
&\leq 32\delta_M n \alpha^2 (1+\epsilon) [KAm\tau]^2 \sum_{k\geq 2} \alpha^{k-2} (1+\epsilon)^{k-2} (KAm\tau)^{k-2} \\
&\leq \frac{32\delta_M n \alpha^2 (1+\epsilon) [KAm\tau]^2}{1-\alpha(1+\epsilon)KAm\tau}, \quad \text{for } 0 < \alpha < \left((1+\epsilon)KAm\tau\right)^{-1}.
\end{aligned}
$$

4.5.5 Proof of Lemma 4.16

Using Assumption 3, we have that $\mathbb{E}_\pi[h_{i,j}] = \int_{x_i} P^{i-j+t_n}(X_{j-t_n}, dx_i) \int_{x_j} h_{i,j}(x_i, x_j) d\pi(x_j)$. Hence we get,

$$
\begin{aligned}
(2a) &:= \left| \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i, X_j)\right] - \mathbb{E}_\pi\left[h_{i,j}\right] \right| \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \left| \int_{x_i} \int_{x_j} h_{i,j}(x_i, x_j) P^{i-j+t_n}(X_{j-t_n}, dx_i) \left(P^{j-i}(x_i, dx_j) - d\pi(x_j)\right) \right| \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \|h_{i,j}\|_\infty \underbrace{\int_{x_i} P^{i-j+t_n}(X_{j-t_n}, dx_i)}_{=1} \underbrace{\sup_y \int_{x_j} \left|P^{j-i}(y, dx_j) - d\pi(x_j)\right|}_{=\sup_y \|P^{j-i}(y,\cdot) - \pi\|_{\mathrm{TV}}} \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \|h_{i,j}\|_\infty L\rho^{j-i} \leq \sum_{j=2}^{n} \sum_{i=(j-t_n+1)\vee 1}^{j-\lfloor \frac{t_n}{2} \rfloor} \|h_{i,j}\|_\infty L\rho^{t_n/2} \leq LAt_n,
\end{aligned}
$$

where we used that $\rho^{t_n/2} \leq \rho^{r\log(n)/2} = n^{r\log(\rho)/2} \leq n^{-1}$. Indeed $1 + r\log(\rho)/2 < 0$ because we choose r such that $r > 2(\log(1/\rho))^{-1}$. With an analogous approach, we bound the term $(2c)$ as follows.

$$
\begin{aligned}
(2c) &:= \left| \sum_{j=2}^{n} \sum_{i=(j-\lfloor \frac{t_n}{2} \rfloor+1)\vee 1}^{j-1} \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i, X_j)\right] - \mathbb{E}\left[h_{i,j}(X_i, X_j)\right] \right| \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-\lfloor \frac{t_n}{2} \rfloor+1)\vee 1}^{j-1} \left| \int_{x_j} \int_{x_i} P^{j-i}(x_i, dx_j) h_{i,j}(x_i, x_j) \left(P^{i-j+t_n}(X_{j-t_n}, dx_i) - d\pi(x_i)\right) \right| \\
&\leq \sum_{j=2}^{n} \sum_{i=(j-\lfloor \frac{t_n}{2} \rfloor+1)\vee 1}^{j-1} \|h_{i,j}\|_\infty \sup_z \|P^{i-j+t_n}(z, \cdot) - \pi\|_{\mathrm{TV}} \\
&\leq \sum_{j=\lfloor \frac{t_n}{2} \rfloor}^{n} \sum_{i=(j-\lfloor \frac{t_n}{2} \rfloor+1)}^{j-1} \|h_{i,j}\|_\infty L\rho^{i-j+t_n} + \sum_{j=2}^{\lfloor \frac{t_n}{2} \rfloor} \sum_{i=(j-\lfloor \frac{t_n}{2} \rfloor+1)\vee 1}^{j-1} \|h_{i,j}\|_\infty L\rho^{i-j+t_n}
\end{aligned}
$$

$$\leq \sum_{j=\lfloor \frac{t_n}{2} \rfloor}^{n} \sum_{i=(j-\lfloor \frac{t_n}{2} \rfloor+1)}^{j-1} \|h_{i,j}\|_\infty L\rho^{t_n/2} + t_n^2 \|h_{i,j}\|_\infty L \leq LA\left(t_n^2 + nt_n\rho^{t_n/2}\right) \leq 2LAt_n^2,$$

where we used that $\rho^{t_n/2} \leq \rho^{r\log(n)/2} = n^{r\log(\rho)/2} \leq n^{-1}$.

4.5.6 Proof of Proposition 4.2

Since the split chain has the same distribution as the original Markov chain, we get that $(\widetilde{X}_i)_i$ is ψ-irreducible for some measure ψ and uniformly ergodic. From [Meyn and Tweedie, 1993, Theorem 16.0.2], Assumption 1 ensures that for every measurable set $A \subset E \times \{0,1\}$ such that $\psi(A) > 0$, there exists some $\kappa_A > 1$ such that

$$\sup_x \mathbb{E}[\kappa_A^{\tau_A} | \widetilde{X}_1 = x] < \infty,$$

where $\tau_A := \inf\{n \geq 1 \ : \ \widetilde{X}_n \in A\}$ is the first hitting time of the set A. Let us recall that T_1 and T_2 are defined as hitting times of the atom of the split chain $E \times \{1\}$ which is accessible (i.e. the atom has a positive ψ-measure). Hence, there exist $C > 0$ and $\kappa > 1$ such that,

$$\sup_x \mathbb{E}[\kappa^{\tau_{E\times\{1\}}} | \widetilde{X}_1 = x] = \sup_x \mathbb{E}[\exp(\tau_{E\times\{1\}} \log(\kappa)) | \widetilde{X}_1 = x] \leq C.$$

Considering $k \geq 1$ such that $C^{1/k} \leq 2$, a straight forward application of Jensen inequality gives that $\max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1}) \leq k/\log(\kappa)$.

Chapter 5

Three rates of convergence or separation via U-statistics in a dependent framework

Chapter Abstract

In this chapter, we rely on the progress we made in Chapter 4 regarding measure concentration in a Markovian framework to push further the current state of knowledge in active areas of research in Probability, Statistics and Machine Learning.
As a first result, we provide for the first time a MCMC procedure to estimate the spectra of signed integral operators in Section 5.3 with theoretical guarantees. In Section 5.4, we give an online-to-batch conversion result for online learning with pairwise loss functions in a Markovian framework. Finally, Section 5.5 contains a multiple testing procedure based on the L^2 distance to test that the density f of the stationary distribution of an observed Markov chain equals some prescribed density f_0.

Chapter Content

5.1 Introduction

5.1.1 Context and Contributions

Our new results - that we referred to as applications for brevity - push further the current state of knowledge in three different active areas of research in Probability, Statistics and Machine Learning. Although the recent progress in concentration inequality for U-statistics with dependent random variables is a key element in our proofs, our contributions are not a direct consequence of it. In this section, we briefly introduce the different research areas in which our work is embedded and present our main contributions.

- **Estimation of spectra of signed integral operator with MCMC algorithms** (Section 5.3)
  We study the convergence of sequence of spectra of kernel matrices towards the spectrum of some integral operator. Previous important works may include Adamczak and Bednorz [2015a] and, as far as we know, they all assume that the kernel is of positive-type (*i.e.*, giving an integral operator with non-negative eigenvalues). For the first time, we prove a non-asymptotic result of convergence of spectra for kernels that are not of positive-type. We further prove that *independent Hastings algorithms* are valid sampling schemes to apply our result.
  In Section 5.3.2, we propose a detailed comparison between our result and the one from Adamczak and Bednorz [2015a]. We explain why working with integral operators of positive-type allows Adamczak and Berdnorz to make use of a powerful decoupling technique. Thanks to this elegant argument, they are back to prove a concentration inequality for a sum of Banach space valued random variables where the i-th summand depends only on the i-th visited state of the Markov chain. By considering signed integral operators, the approach of the former paper cannot be adapted. As a result, our proof relies on completely different arguments which are highlighted in Section 5.3.2.

- **Online learning with pairwise loss functions** (Section 5.4)
  In Machine Learning, several important problems involve a pairwise loss function, i.e. a loss function which depends on a pair of examples. One typical example is the problem of metric learning [cf. Jin et al., 2009] where one aims to learn a metric so that instances with the same labels are close while ones with different labels are far away from each other. Other pairwise learning tasks include preference learning Xing et al. [2002], ranking Agarwal and Duchi [2012], gradient learning Meir and Zhang [2003] and AUC maximization Zhao et al. [2011]. Batch learning algorithms with pairwise loss functions have been extensively studied and their generalization properties have been well established. However, batch algorithms have some limitations especially when data becomes available in a sequential order or for large scale learning problems where their computational cost can be prohibitive. Online algorithms have been designed to efficiently solve learning problems in such situations: they deal with data coming on fly and try to improve the learned model along time based on the new observations. The performance of online learning algorithms is typically analyzed through the notion of *regret* which compares the payoff obtained by the algorithm along time with the one that would have been obtained by taking the optimal decision at each time step [cf. Bubeck and Cesa-Bianchi, 2012]. The regret quantifies the number of mistakes made by the algorithm without requiring assumptions on the way the training sequence is generated. When the sequence of observations is the realization of some stochastic process, one can analyze online algorithms through a different lens by wondering how they generalize on future examples. More precisely, we would like to convert a regret bound of an online learner into a control of the excess risk. In the online learning research community, these types of results are called *online-to-batch conversion* and we refer to [Hoi et al., 2021, Section 3.7] for a comprehensive introduction to this topic. Online-to-batch conversion results for online learning with univariate or pairwise loss functions working with i.i.d. samples have been considered for quite a while in both Machine Learning and Statistics literature [cf. Chen and Lei, 2018, Guo et al., 2017, Wang et al., 2012, Ying and Zhou, 2017]. For dependent data sequences, generalization bounds for online algorithms have also been proved in the last decades with univariate loss functions [cf. Agarwal and Duchi, 2012]. However, theoretical guarantees for the generalization performance of online algorithms with pairwise loss functions with non i.i.d. data have been so far understudied. Our work is one of the first to bring results regarding this problem. In Section 5.4.1.4, we establish clear connections with the existing literature.

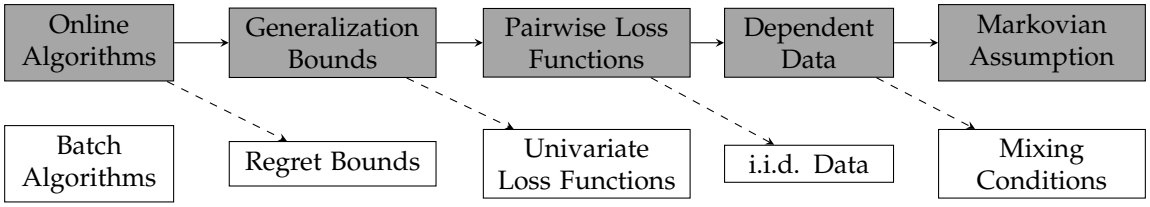


Figure 5.1: Positioning our contribution in the existing literature for the analysis of online algorithms.

The structure of our proof relies mainly on the work from Wang et al. [2012] where the observations were assumed to be i.i.d. Nevertheless, working with a Markovian dynamic brings extra technicalities that we handle using properties of uniformly ergodic Markov chains, concentration inequalities for U-statistics (of order one and two) of dependent random variables and reversibility of Markov chains by considering the time-reversed sequence. Using the marker ☡, we shed light in Section B on the specific parts of the proof where the arguments used in the i.i.d. framework fail, requiring a specific theoretical work handling a sequence of dependent observations.

- **Adaptive goodness-of-fit tests in a density model** (Section 5.5)
  Several works have already proposed goodness-of-fit tests for the density of the stationary distribution of a sequence of dependent random variables. In Li and Tkacz [2001], a test based on an L^2-type distance between the nonparametrically estimated conditional density and its model-based parametric counterpart is proposed. In Bai [2003] a Kolmogorov-type test is considered. Chwialkowski et al. [2016] derive a test procedure for τ-mixing sequences using Stein discrepancy computed in a reproducing kernel Hilbert space. In all the above mentioned papers, asymptotic properties of the test statistic are derived but no non-asymptotic analysis of the methods is conducted. As far as we know, we are the first to provide a non-asymptotic condition on the classes of alternatives ensuring that the statistical test reaches a prescribed power working in a dependent framework.

5.1.2 Outline

In Section 5.2, we introduce useful notations for this section and we present the concentration inequality for U-statistics from Chapter 4 that is an important argument of our proofs. The next three sections are dedicated to our main results. We start by providing a convergence result for the estimation of spectra of integral operators with MCMC algorithms (see Section 5.3). We show that independent Hastings algorithms satisfy under mild conditions the assumptions of Section 5.2.2 and we illustrate our result with the estimation of the spectra of some Mercer kernels. For the second application of our concentration inequality, we investigate the generalization performance of online algorithms with pairwise loss functions in a Markovian framework (see Section 5.4). We motivate the study of such problems and we provide an online-to-batch conversion result. In a third and final application, we propose a goodness-of-fit test for the density of the stationary measure of a Markov chain (see Section 5.5). We give an explicit condition on the set of alternatives to ensure that the statistical test proposed reaches a prescribed power. The proofs related to the three applications are given in Section 5.6, Section 5.7 and Sections 5.8 respectively.

5.2 Notations and Preliminaries

5.2.1 Notations

Let us consider an arbitrary measurable space $(F, \mathcal{F})$. For any measure ω on $(F, \mathcal{F})$, the total variation norm of ω is defined by $\|\omega\|_{\mathrm{TV}} := \sup_{A\in\mathcal{F}} |\omega(A)|$. The space of square summable functions on F with respect to the measure ω is

$$L^2(\omega) := \{f : F \to \mathbb{R} \text{ measurable } | \int_F f(x)^2 d\omega(x) < \infty\}.$$

Note that when $F = \mathbb{R}$ and ω is the Lebesgue on $\mathbb{R}$, we will denote $L^2(\omega) = L^2(\mathbb{R})$. Endowed with the inner product $(f, g) \in L^2(\omega) \times L^2(\omega) \mapsto \langle f, g\rangle := \int_F f(x)g(x)d\omega(x)$, $L^2(\omega)$ is a Hilbert space and

we denote by $\|\cdot\|_2$ the norm induced by $\langle\cdot,\cdot\rangle$. For any function $h : F \to \mathbb{R}$, we define the supremum norm of h by $\|h\|_\infty := \sup_{x\in F} |h(x)|$. We further denote $\mathbb{N}^* := \mathbb{N}\backslash\{0\}$ and $\mathcal{B}(\mathbb{R})$ the Borel algebra on $\mathbb{R}$. For any $x \in \mathbb{R}_+$, we denote by $\lfloor x\rfloor$ (resp. $\lceil x\rceil$) the largest integer that is less than or equal to x (resp. the smallest integer greater than or equal to x). For any $x, y \in \mathbb{R}$, we set $x \vee y := \max(x, y)$ and $x \wedge y := \min(x, y)$. For any integers i, j, the Kronecker symbol $\delta_{i,j}$ is equal to 1 if $i = j$ and is equal to 0 otherwise.

5.2.2 Concentration inequality for U-statistics of uniformly ergodic Markov chains

In this section, we propose a brief reminder of the concentration inequality from Chapter 4 for U-statistics of uniformly ergodic discrete time Markov chains that will be an essential tool in our proofs. Let us mention that we do not work with the concentration inequality from Shen et al. [2020] since it only holds for stationary chains if the kernel h is π-canonical (see Assumption 3). Stationarity may be seen as a really strong assumption which would make our main results from Section 5.3 of little interest since MCMC methods are used when we are not able to directly sample from the distribution π. Regarding Sections 5.4 and 5.5, the concentration inequality for U-statistics used in the proofs of our results needs to hold for any initial distribution of the chain.

Let (E, Σ) be a measurable space. We consider a Markov chain $(X_i)_{i\geq 1}$ on (E, Σ) with transition kernel $P : E \times E \to [0, 1]$ and with a unique stationary distribution π. We consider some measurable function $h : (E \times E, \Sigma \otimes \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and we are interested in the following U-statistic

$$U_{\text{stat}}(n) := \sum_{1\leq i\neq j\leq n} \big(h(X_i, X_j) - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}[h(X, Y)]\big) .$$

We will work under the following set of assumptions.

Assumption 1. *The Markov chain $(X_i)_{i\geq 1}$ is ψ-irreducible for some maximal irreducibility measure ψ on Σ [cf. Meyn and Tweedie, 1993, Section 4.2]. Moreover, there exist some natural number m and a constant $\delta_m > 0$ such that*

$$\forall x \in E,\ \forall A \in \Sigma, \quad \delta_m \mu(A) \leq P^m(x, A). \tag{5.1}$$

for some probability measure μ.

A Markov chain satisfying Assumption 1 is called uniformly ergodic and admits a unique stationary distribution denoted by π. Assumption 1 also implies that the regeneration times associated to the split chain are exponentially integrable, meaning that their Orlicz norm with respect to the function $\psi_1(x) = \exp(x) - 1$ are bounded by some constant $\tau > 0$. We refer to Section 4.2.3 in Chapter 4 for details.

Assumption 2 can be read as a reverse Doeblin's condition and is used in Chapter 4 as a decoupling tool. Let us recall that we give in Chapter 4 several natural examples for which this condition holds.

Assumption 2. *There exist $\delta_M > 0$ and some probability measure ν such that*

$$\forall x \in E,\ \forall A \in \Sigma, \quad P(x, A) \leq \delta_M \nu(A).$$

The last assumption introduces the notion of π-*canonical* kernel, which is the counterpart in the Markovian setting of the canonical (or degenerate) property of the independent framework.

Assumption 3. *Denoting by π the stationary distribution of the Markov chain $(X_i)_{i\geq 1}$, we assume that $h : (E \times E, \Sigma \otimes \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable, bounded and is π-canonical, namely*

$$\forall x, y \in E, \quad \mathbb{E}_\pi[h(X, x)] = \mathbb{E}_\pi[h(X, y)] = \mathbb{E}_\pi[h(x, X)] = \mathbb{E}_\pi[h(y, X)].$$

This common expectation will be denoted by $\mathbb{E}_\pi[h]$.

Let us mention that several important kernels are π-canonical. This is the case of translation-invariant kernels which have been widely studied in the Machine Learning community [cf. Lerasle et al., 2016]. Another example of π-canonical kernel is a rotation invariant kernel when $E = \mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ with π also rotation invariant (see Chapter 3). Note also that if the kernel h is not π-canonical, the U-statistic decomposes into a linear term and a π-canonical U-statistic. This is called the *Hoeffding*

decomposition [cf. Giné and Nickl, 2016, p.176] and takes the following form

$$\sum_{i\neq j}\left(h(X_i,X_j)-\mathbb{E}_{(X,Y)\sim\pi\otimes\pi}[h(X,Y)]\right)=\sum_{i\neq j}\left(\widetilde{h}(X_i,X_j)-\mathbb{E}_{\pi}\left[\widetilde{h}(X,\cdot)\right]\right)$$
$$+\sum_{i\neq j}\left(\mathbb{E}_{X\sim\pi}\left[h(X,X_j)\right]-\mathbb{E}_{(X,Y)\sim\pi\otimes\pi}\left[h(X,Y)\right]\right)$$
$$+\sum_{i\neq j}\left(\mathbb{E}_{X\sim\pi}\left[h(X_i,X)\right]-\mathbb{E}_{(X,Y)\sim\pi\otimes\pi}\left[h(X,Y)\right]\right),$$

where the kernel $\widetilde{h}$ is π-canonical with

$$\forall x,y\in E,\quad \widetilde{h}(x,y)=h(x,y)-\mathbb{E}_{X\sim\pi}\left[h(x,X)\right]-\mathbb{E}_{X\sim\pi}\left[h(X,y)\right].$$

We will use this method several times in our proofs (for example in Eq.(5.19)).
We are now ready to remind one of the main result from Chapter 4 that is one key theoretical tool to derive our three contributions presented in the next sections.

Theorem 5.1. *Suppose that Assumptions 1, 2 and 3 are satisfied. Then there exist constants $\beta,\kappa>0$ (depending on the Markov chain $(X_i)_{i\geq 1}$) such that for any $u\geq 1$ and any $n\geq 2$, with probability at least $1-\beta e^{-u}\log n$,*

$$\frac{2}{n(n-1)}U_{\mathrm{stat}}(n)\leq\kappa\|h\|_{\infty}\log n\left\{\frac{u}{n}+\left(\frac{u}{n}\right)^2\right\}.$$

5.3 Estimation of spectra of signed integral operator with MCMC algorithms

5.3.1 MCMC estimation of spectra of signed integral operators

Let us consider a Markov chain $(X_n)_{n\geq 1}$ on E satisfying the assumptions of Theorem 5.1 with stationary distribution π, and some symmetric kernel $h:E\times E\to\mathbb{R}$ such that $h\in L^2(\pi\otimes\pi)$. We can associate to h the kernel of a linear operator $\mathbf{H}$ defined by

$$\mathbf{H}f(x):=\int_E h(x,y)f(y)d\pi(y). \tag{5.2}$$

This is a Hilbert-Schmidt operator on $L^2(\pi)$ and thus it has a real spectrum consisting of a square summable sequence of eigenvalues [cf. Conway, 2019, p.267]. In the following, we will denote the eigenvalues of $\mathbf{H}$ by $\lambda(\mathbf{H}):=(\lambda_1,\lambda_2,\dots)$. For some $n\in\mathbb{N}^*$, we consider

$$\widetilde{\mathbf{H}}_n:=\frac{1}{n}\left(h(X_i,X_j)\right)_{1\leq i,j\leq n}\text{ and }\mathbf{H}_n:=\frac{1}{n}\left((1-\delta_{i,j})h(X_i,X_j)\right)_{1\leq i,j\leq n}, \tag{5.3}$$

with respective eigenvalues $\lambda(\widetilde{\mathbf{H}}_n)$ and $\lambda(\mathbf{H}_n)$. Following [Koltchinskii and Giné, 2000, Section 2], we introduce in Definition 5.2 the rearrangement distance δ_2 which measures closeness of spectra.

Definition 5.2. Given two sequences x,y of reals – completing finite sequences by zeros – such that

$$\sum_i x_i^2+y_i^2<\infty,$$

we define the ℓ_2 rearrangement distance $\delta_2(x,y)$ as

$$\delta_2^2(x,y):=\inf_{\sigma\in\mathfrak{S}}\sum_i(x_i-y_{\sigma(i)})^2,$$

where $\mathfrak{S}$ is the set of permutations of natural numbers. δ_2 is a pseudometric on ℓ_2, where ℓ_2 is the Hilbert space of all square summable sequences.

Theorem 5.3 gives conditions ensuring that both the spectrum of $\mathbf{H}_n$ and the one of $\widetilde{\mathbf{H}}_n$ converge towards the spectrum of the integral operator $\mathbf{H}$ as $n\to\infty$. Theorem 5.3 holds under Assumption 5 that

we discuss in details in Section 5.3.2. The proof of Theorem 5.3 is postponed to Section 5.6.

Assumption 5. *$h : E \times E \to \mathbb{R}$ is a bounded and symmetric function. Moreover there exist continuous functions $\phi_r : E \to \mathbb{R}$, $r \in I$ (where $I = \mathbb{N}$ or $I = 1, \dots, N$) that form an orthonormal basis of $L^2(\pi)$ and a sequence of real numbers $(\lambda_r)_{r\in I} \in \ell_2$ such that we have pointwise*

$$h(x, y) = \sum_{r\in I} \lambda_r \phi_r(x)\phi_r(y),$$

with $\Upsilon := \sup_{r\in I} \|\phi_r\|_\infty^2 < \infty$ and $S := \sup_{x\in E} \sum_{r\in I} |\lambda_r|\phi_r(x)^2 < \infty$.
We further denote $\Lambda := \sup_{r\in I} |\lambda_r|$.

Theorem 5.3. *Let $(X_i)_{i\geq 1}$ be a Markov chain on E satisfying Assumptions 1 and 2 described in Section 5.2.2 with stationary distribution π. Suppose that Assumption 5 is satisfied. Then for any $t > 0$,*

$$\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 \geq \frac{S^2(1+\kappa)\log n}{n} + 2\sum_{i>\lceil n^{1/4}\rceil, i\in I} \lambda_i^2 + t\right)$$
$$\leq 32\sqrt{n}\exp\left(-\mathcal{C}\min\left(nt^2, \sqrt{n}t\right)\right) + \beta\log(n)\exp\left(-\frac{n}{\log n}\min\left(\mathcal{B}t, (\mathcal{B}t)^{1/2}\right)\right).$$

where for some universal constant $K > 0$, we have $\mathcal{B} = (K\kappa S)^{-1}$, $\mathcal{C} = \left(K^{1/2}m\tau(S + \Lambda\Upsilon)\right)^{-2}$. $\kappa > 0$ and $\beta > 0$ are the constants from Theorem 5.1 and depend on the Markov chain. We refer to Assumption 1 and the following remark for the definitions of the constants m and τ.

Remark. The same bound holds for $\delta_2(\lambda(\mathbf{H}), \lambda(\widetilde{\mathbf{H}}_n))^2$.

5.3.2 Comparison with the existing literature

Known result for *positive* kernels. In Adamczak and Bednorz [2015a], the authors studied the convergence properties of MCMC methods to estimate the spectrum of integral operators with bounded *positive* kernels (*i.e.*, such that $\mathbf{H}$ has non-negative eigenvalues). They show a sub-exponential tail behavior for the δ_2 distance between the spectrum of $\mathbf{H}$ and the one of the random matrix $\mathbf{H}_n$. Their result has the merit to hold for geometrically ergodic Markov chains, but they work with the restrictive assumption that the eigenvalues of $\mathbf{H}$ are non-negative.
Hence, working with stronger conditions on the Markov chain $(X_i)_{i\geq 1}$, Theorem 5.3 proves a new concentration inequality for the δ_2 distance between $\lambda(\mathbf{H})$ and $\lambda(\mathbf{H}_n)$ that holds for **arbitrary signs of the eigenvalues** of $\mathbf{H}$. Note that the set of operators handled by Theorem 5.3 is not a superset of the ones handled by [Adamczak and Bednorz, 2015a, Theorem 2.2]. The difference lies in the fact that we ask the family of functions $(\phi_r)_{r\in I}$ to be uniformly bounded (cf. Assumption 5). Let us mention that the set of assumptions considered in Adamczak and Bednorz [2015a] implies that $S < \infty$. In the following, we comment our extra assumption $\Upsilon < \infty$ with more details.

1. The basis functions $(\phi_r)_{r\in I}$ are continuous and Assumption 2 typically holds for a compact space E. Hence, by considering that E is compact and that the sequence $(\lambda_r)_{r\in I}$ has finite support (i.e. $I = [N]$ for some natural number N), it holds $\Upsilon < \infty$.

2. Asking for $\Upsilon < \infty$ is only useful to apply a concentration inequality for Markov chains at one specific step of our proof (cf. Eq.(5.16)). Hence this assumption might be weakened by applying other exponential tail control for empirical processes of Markov chains. Nevertheless we point out that Theorem 5.3 holds for uniformly ergodic Markov chains which is equivalent to the standard *drift condition* where the drift function V is bounded [cf. Meyn and Tweedie, 1993, Chap.16]. Hence, the assumptions needed for the exponential inequality from [Adamczak and Bednorz, 2015b, Theorem 1.1] or the one from [Durmus et al., 2021, Theorem 5] imply that $\Upsilon < \infty$. Hence, weakening the condition $\Upsilon < \infty$ seems challenging but we believe that it might be done in some specific settings using for instance the work from [Ciolek and Bertail, 2019, Section 3.2].

Proof structures. In the following, we describe the proof structure of [Adamczak and Bednorz, 2015a, Theorem 2.2], allowing the reader to understand why their approach can only handle positive kernel.

Considering $f : E \to L^2(\pi)$ defined by

$$\forall x \in E, \quad f(x) := \sum_{r \in I} \sqrt{\lambda_i} \phi_r(x) \phi_r,$$

it holds

$$\mathbb{E}_{X \sim \pi} \left[f(X) \otimes f(X) \right] = \mathbf{H},$$

where $\otimes$ denotes the tensor product and where the expectation on the left-hand side is the Bochner integral in the Hilbert space of Hilbert–Schmidt operators. The empirical counterpart of the operator $\mathbf{H}$ can thus be written as the sum of independent random rank one operators and is given by

$$\mathbf{K}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i) \otimes f(X_i),$$

i.e. for any $u \in L^2(\pi)$, $\mathbf{K}_n u := \frac{1}{n} \sum_{i=1}^{n} \langle f(X_i), u \rangle f(X_i)$.
$\mathbf{K}_n$ can be written as $A_n A_n^\top$ where $A_n : \mathbb{R}^n \to L^2(\pi)$ is defined by $A_n e_i = n^{-1/2} f(X_i)$ for all $i \geq 1$ ($e_1, \ldots, e_n$ being the standard basis in $\mathbb{R}^n$). It is straightforward to show that $\forall i, j \in [n]$, $\langle A_n^\top A_n e_i \, , \, e_j \rangle = \frac{1}{n} h(X_i, X_j)$ so that $\widetilde{\mathbf{H}}_n = A_n^\top A_n$. Hence, $\lambda(\widetilde{\mathbf{H}}_n) = \lambda(\mathbf{K}_n)$, which leads to

$$\begin{aligned} \delta_2 \left(\lambda(\mathbf{H}), \lambda(\widetilde{\mathbf{H}}_n) \right) &= \delta_2 \left(\lambda(\mathbf{H}), \lambda(\mathbf{K}_n) \right) \\ &\leq \| \frac{1}{n} \sum_{i=1}^{n} f(X_i) \otimes f(X_i) - \mathbb{E}[f(X) \otimes f(X)] \|_2, \end{aligned} \tag{5.4}$$

where the last inequality follows from the infinite-dimensional version of the Hoffman–Wielandt inequality proved by Bhatia and Elsner [1994]. The right hand side of Eq.(5.4) shows that we are back to deal with a sum of Banach space valued functions of the Markov chain $(X_i)_{i \geq 1}$. After using the standard splitting technique [cf. Meyn and Tweedie, 1993, Section 5.1], the proof is concluded by applying Bernstein-type inequality for sums of independent Banach space valued random variables. Assuming that the eigenvalues $(\lambda_r)_{r \in I}$ are non-negative should thus be understood as **a sufficient condition for decoupling** in the sense that the i-th summand in the right hand side of Eq.(5.4) only depends on X_i. As a consequence, by considering signed integral operators, we cannot adapt the proof proposed by Adamczak and Bednorz [2015a] and we followed a completely different approach to prove Theorem 5.3.
Let us now outline the principal steps of the proof of Theorem 5.3. For any natural number R, we denote $\mathbf{H}^R$ the integral operator with kernel function h_R at resolution R, namely

$$h_R(x, y) := \sum_{r \in I, r \leq R} \lambda_r \phi_r(x) \phi_r(y), \quad \mathbf{H}^R f(x) := \int_E h_R(x, y) f(y) d\pi(y).$$

We define $\widetilde{\mathbf{H}}_n^R$ and $\mathbf{H}_n^R$ analogously by using the kernel h_R in Eq.(5.3). Using the triangle inequality, we split the distance $\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))$ into four terms.

1. $\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}^R))$ is a bias term induced by working at resolution R.

2. A non-trivial preliminary work allows to prove that $\delta_2(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R))$ can be written as an empirical process of the Markov chain $(X_i)_{i \geq 1}$ whose tail can be controlled by applying concentration inequalities for sums of functions of uniformly ergodic Markov chains (this is where we use the assumption that Υ is finite).

3. Since the matrices $\mathbf{H}_n^R$ and $\widetilde{\mathbf{H}}_n^R$ only differ at diagonal elements, $\delta_2(\lambda(\widetilde{\mathbf{H}}_n^R), \lambda(\mathbf{H}_n^R))$ can be coarsely bounded by $n^{-1/2} \| h_R \|_\infty$.

4. Applying the Hoffman-Wielandt inequality, one can notice that $\delta_2(\lambda(\mathbf{H}_n^R), \lambda(\mathbf{H}_n))$ can be upper-bounded by a U-statistic of order two of the Markov chain $(X_i)_{i \geq 1}$. We control the tail behaviour of this U-statistic by applying Theorem 5.1.

The proof is then concluded by choosing the resolution level R so that $R^2 = \lceil \sqrt{n} \rceil$.

5.3.3 Admissible sampling schemes: Independent Hastings algorithm

One can use the previous result to estimate the spectrum of the integral operator $\mathbf{H}$ using MCMC methods. To do so, we need to make sure that the Markov chain used for the MCMC method satisfies the conditions of Theorem 5.1. It is for example well known that Metropolis random walks on $\mathbb{R}$ are not uniformly ergodic [cf. Meyn and Tweedie, 1993]. In the following, we show that an independent Hastings algorithm can be used on bounded state space to generate a uniformly ergodic chain with the desired stationary distribution.

5.3.3.1 Independent Hastings algorithm on bounded state space.

Let $E \subset \mathbb{R}^k$ be a bounded subset of $\mathbb{R}^k$ equipped with the Borel σ-algebra $\mathcal{B}(E)$. Denoting λ_{Leb} the Lebesgue measure on E, we consider a measure π on E - which is only known up to a factor - with density f_π with respect to λ_{Leb}, and a probability density q with respect to λ_{Leb}, satisfying $f_\pi(y), q(y) > 0$ for all $y \in E$. In the independent Hastings algorithm, a candidate transition generated according to the law $q\lambda_{Leb}$ is then accepted with probability $\alpha(x, y)$ given by

$$\alpha(x, y) = \min\left(1, \frac{f_\pi(y)q(x)}{f_\pi(x)q(y)}\right).$$

With an approach similar to Theorem 2.1 from Mengersen and Tweedie [1996], Proposition 5.4 shows that under some conditions on the densities f_π and q, the independent Hastings algorithm satisfies the Assumptions 1 and 2.

Proposition 5.4. *Let us assume that* $\sup_{x\in E} q(x) < \infty$ *and that there exists* $\beta > 0$ *such that*

$$\frac{q(y)}{f_\pi(y)} > \beta, \quad \forall y \in E.$$

Then, the independent Hastings algorithm satisfies the Assumptions 1 and 2.

Proof of Proposition 5.4. We denote by P the transition kernel of the Markov chain generated with the independent Hastings algorithm. For any $x \in E$, the density with respect to λ_{Leb} of the absolutely continuous part of $P(x, dy)$ is $p(x, \cdot) = q(\cdot)\alpha(x, \cdot)$, while the singular part is given by $\mathbb{1}_x(\cdot)\left(\int_{z\in E} q(z)\alpha(x, z)d\lambda_{Leb}(z)\right)$. For fixed $x \in E$, we have either $\alpha(x, y) = 1$ in which case $p(x, y) = q(y) \geq \beta f_\pi(y)$, or else

$$p(x, y) = q(y)\frac{f_\pi(y)q(x)}{f_\pi(x)q(y)} = q(x)\frac{f_\pi(y)}{f_\pi(x)} \geq \beta f_\pi(y).$$

We deduce that for any $x \in E$, it holds

$$P(x, A) \geq \beta \int_{y\in A} f_\pi(y)d\lambda_{Leb}(y),$$

which proves that the chain is uniformly ergodic (cf. Eq.(5.1)). Hence Assumption 1 is satisfied. Assumption 2 trivially holds since E is bounded and $\sup_{y\in E} q(y) < \infty$. □

From Proposition 5.4 and Theorem 5.3, we deduce that one can use a MCMC approach to estimate the spectrum of a signed integral operator $\mathbf{H}$ (that satisfies Assumption 5) as defined in Eq.(5.2) where E is a bounded subset of $\mathbb{R}^k$. More precisely, if the density f_π of Eq.(5.2) is known up to a factor and if there exists some probability density q with respect to λ_{Leb} satisfying the assumptions of Proposition 5.4, the Independent Hastings algorithm provides a Markov chain that satisfies Assumptions 1 and 2. Hence the non-asymptotic bound from Theorem 5.3 holds. We put this methodology into action in the new section by estimating the spectrum of some Mercer kernels on the d-dimensional sphere.

5.3.4 Estimation of the spectrum of Mercer kernels

In this example, we illustrate Theorem 5.3 by computing the eigenvalues of an integral operator naturally associated with a Mercer kernel using a MCMC algorithm. A function $h : E \times E \to \mathbb{R}$ is called a

Mercer kernel if E is a compact metric space and $h : E \times E \to \mathbb{R}$ is a continuous symmetric and positive definite function. It is well known that if h is a Mercer kernel, then the integral operator L_h associated with h is a compact and bounded linear operator, self-adjoint and semi-definite positive. The spectral theorem implies that if h is a Mercer kernel, then there is a complete orthonormal system $(\phi_1, \phi_2, \dots)$ of eigenvectors of L_h. The eigenvalues $(\lambda_1, \lambda_2, \dots)$ are real and non-negative. The Mercer Theorem [see e.g. Christmann and Steinwart, 2008, Theorem 4.49] shows that the eigen-structure of L_h can be used to get a representation of the Mercer kernel h as a sum of a convergent sequence of product functions for the uniform norm. In this context, Theorem 5.3 allows to derive the convergence rate in the δ_2 metric of the estimated spectrum towards the one of the integral operator $\mathbf{H}$ as presented in Proposition 5.5.

Proposition 5.5. *We keep the notations and the assumptions of Theorem 5.3. We assume further that there exists $s > 0$, a (Sobolev) regularity parameter, such that for some constant $C(s) > 0$,*

$$\forall R > 1, \quad \sum_{i>R} \lambda_i^2 \leq C(s) R^{-2s}.$$

Then it holds

$$\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 = \begin{cases} \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{n}}\right) & \text{if } s \geq 1 \\ \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n^{s/2}}\right) & \text{if } s \in (0,1) \end{cases}.$$

Proof of Proposition 5.5. Proposition 5.5 directly follows from Theorem 5.3 by choosing $t = \sqrt{\frac{\log n}{n}}$. □

To illustrate our purpose, we consider the d-dimensional sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. We consider a positive definite kernel on $\mathbb{S}^{d-1}$ defined by $\forall x, y \in \mathbb{S}^{d-1}, \quad h(x,y) = \psi(x^\top y)$ where $\psi : [-1,1] \to \mathbb{R}$ is continuous. From the Funk-Hecke Theorem [see e.g. Müller, 2012, p.30], we know that the eigenvalues of the Mercer kernel h are

$$\lambda_k = \lambda_{Leb}(\mathbb{S}^{d-2}) \int_{-1}^{1} \psi(t) P_k(d;t) \left(1-t^2\right)^{\frac{d-3}{2}} dt, \tag{5.5}$$

where $\lambda_{Leb}(\mathbb{S}^{d-2})$ is the surface area of $\mathbb{S}^{d-2}$ and $P_k(d;t)$ is the Legendre polynomial of degree k in dimension d. For any $k \in \mathbb{N}$, the multiplicity of the eigenvalue λ_k is the dimension of the space of spherical harmonics of degree k. To build the Markov chain $(X_i)_{i\geq 1}$, we start by sampling randomly X_1 on $\mathbb{S}^{d-1}$. Then, for any $i \in \{2, \dots, n\}$, we sample

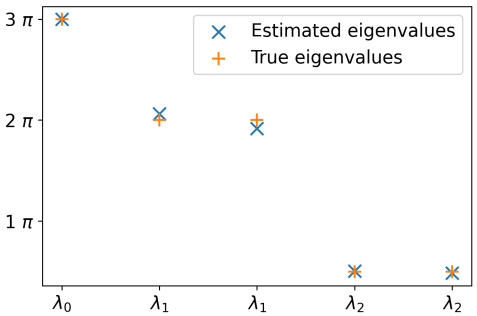
- a unit vector $Y_i \in \mathbb{S}^{d-1}$ uniformly, orthogonal to X_{i-1},
- a real $r_i \in [-1,1]$ encoding the distance between X_{i-1} and X_i. r_i is sampled from a distribution $f_{\mathcal{L}} : [-1,1] \to [0,1]$.

then X_i is defined by

$$X_i = r_i \times X_{i-1} + \sqrt{1-r_i^2} \times Y_i \,.$$

By assuming that $\min_{r\in[-1,1]} f_{\mathcal{L}}(r) > 0$ and $\|f_{\mathcal{L}}\|_\infty < \infty$, Assumptions 1 and 2 hold and the stationary distribution of the chain $(X_i)_{i\geq 1}$ is the Haar measure on $\mathbb{S}^{d-1}$ (cf. Chapter 3).
In Figure 5.2, we plot the non-zero eigenvalues using function $\psi : t \mapsto (1+t)^2$ and taking $f_{\mathcal{L}}$ proportional to $r \mapsto f_{(5,1)}(\frac{r+2}{4})$ where $f_{(5,1)}$ is the pdf of the Beta distribution with parameter $(5,1)$. We plot both the true eigenvalues and the ones computed using a MCMC approach.

Figure 5.2: Consider function $\psi : t \mapsto (1+t)^2$, $d = 2$ and $n = 1000$. The true eigenvalues can be computed using (5.5), but in this case, we know the exact values of the three non-zero eigenvalues namely $\lambda_0 = 3\pi$, $\lambda_1 = 2\pi$ and $\lambda_2 = \pi/2$. Their respective multiplicities are 1, 2 and 2. The estimated eigenvalues are the eigenvalues of the matrix $\mathbf{H}_n = \frac{1}{n}\left((1-\delta_{i,j})\psi(X_i^\top X_j)\right)_{1\leq i,j\leq n}$ where the n points $X_1, X_2, \dots, X_n$ are sampled on the Euclidean sphere $\mathbb{S}^{d-1}$ using a Markovian dynamic.



5.3.5 Some perspectives

In this section, we shed light on some research fields where our results may find an echo. More precisely, we broaden our horizons to show the practical importance for many learning algorithms to estimate the eigenvalues and/or the eigenvectors of data-dependent matrices. This is for example the case for Principal Component Analysis (PCA) or some manifold methods [cf. Rosasco et al., 2010]. It appears that these matrices can often be interpreted as the empirical versions of continuous objects such as integral operators. As highlighted in Rosasco et al. [2010], the theoretical analysis of the above mentioned learning algorithms requires to quantify the difference between the eigen-structure of the empirical operators and their continuous counterparts. In the following, we focus on two examples: the estimation of the entire spectrum of a Markov operator and the generalization performance of neural networks

Estimation of the entire spectrum of a Markov operator. In Chakraborty and Khare [2019], the authors recognize that the literature to estimate the entire spectrum of a Markov operator is rather sparse. They provide in their paper several important practical applications of the estimation of the whole spectrum of a Markov operator, such as the computation of the expected chi-square distance between the stationary distribution $\pi = f\mathrm{d}\nu$ (for some measure ν) of a Markov chain $(X_i)_{i\geq 1}$ and the distribution of X_m. More precisely, denoting by $P^m(\cdot,\cdot)$ the m-step transition density of the Markov chain, the chi-square distance to stationarity after m steps, starting at state x is defined as

$$\chi^2_x(m) := \int \frac{|P^m(x,x') - f(x')|^2}{f(x')}\mathrm{d}\nu(x').$$

If the integral operator associated to the kernel function P is assumed to be Hilbert-Schmidt, then $\chi^2_x(m) = \sum_{i\geq 1}\lambda_i^{2m}\phi_i(x)^2$ where $(\phi_i)_{i\geq 1}$ is an orthonormal basis of $L^2(\pi)$ [cf. Diaconis et al., 2008]. The average or expected chi-square distance to stationarity after m steps is therefore $\sum_{i\geq 1}\lambda_i^{2m}$.
The authors of Chakraborty and Khare [2019] adapt the MCMC algorithm from Adamczak and Bednorz [2015a] to estimate the entire spectrum of a Markov kernel operator that arises in Data Augmentation methods.

The spectral bias of neural networks. Integral operators and their eigenstructures appeared to be essential objects for the theoretical analysis of neural networks (NNs). It is now well-known that over-parametrized neural network have good generalization performance on important learning problems [cf. Zhang et al., 2021]. This fact has been so far explained by noticing empirically that NNs are biased towards learning less complex functions: a phenomenon known as the *spectral bias* [cf. Rahaman et al., 2019]. In Su and Yang [2019], the authors prove that the training process of NNs can be decomposed along different directions defined by the eigenfunctions of some integral operator where each direction has its own convergence rate and the rate is determined by the corresponding eigenvalue.

5.4 Online Learning with Pairwise Loss Functions

5.4.1 Brief introduction to online learning and motivations

5.4.1.1 Presentation of the traditional online learning setting

Online learning is an active field of research in Machine Learning in which data becomes available in a sequential order and is used to update the best predictor for future data at each step. This method aims at learning some function $f : E \to \mathcal{Y}$ where E is the space of inputs and $\mathcal{Y}$ is the space of outputs. At each time step t, we observe a new example $(x_t, y_t) \in E \times \mathcal{Y}$. Traditionally, the random variables (x_t, y_t) are supposed i.i.d. with common joint probability distribution $(x, y) \mapsto p(x, y)$ on $E \times \mathcal{Y}$. In this setting, the loss function is given as $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such that $\ell(f(x), y)$ measures the difference between the predicted value $f(x)$ and true value y. The goal is to select at each time step t a function $h_t : E \to \mathcal{Y}$ in a fixed set $\mathcal{H}$ based on the observed examples until time t (namely $(x_i, y_i)_{1\leq i\leq t}$) such that h_t has "small" risk $\mathcal{R}$ defined by

$$\mathcal{R}(h) = \mathbb{E}_{(X,Y)\sim p}\big[\ell(h(X), Y)\big],$$

where h is any measurable mapping from E to $\mathcal{Y}$.
Online learning is used when data is coming *on the fly* and we do not want to wait for the acquisition of

the complete dataset to take a decision. In such cases, online learning algorithms allow to dynamically adapt to new patterns in the data.

5.4.1.2 Online learning with pairwise loss functions

In some cases, the framework provided in the previous paragraph is not appropriated to solve the task at stake. Consider the example of ranking problems. The state space is E and there exists a function $f : E \to \mathbb{R}$ which assigns to each state $x \in E$ a label $f(x) \in \mathbb{R}$. f naturally defines a partial order on E. At each time step t, we observe an example $x_t \in E$ together with its label $f(x_t)$ and we suppose that the random variables $(x_t)_t$ are i.i.d. with common distribution p. Our goal is to learn the partial order of the items in E induced by the function f. More precisely, we consider a space $\mathcal{H} \subset \{h : E \times E \to \mathbb{R}\}$, called the set of hypotheses. An *ideal* hypothesis $h \in \mathcal{H}$ would satisfy

$$\forall x, u \in E, \quad f(x) \geq f(u) \Leftrightarrow (h(x,u) \geq 0 \text{ and } h(u,x) \leq 0)\,.$$

We consider a loss function $\ell : \mathcal{H} \times E \times E \to \mathbb{R}$ such that $\ell(h,x,u)$ measures the ranking error induced by h and a typical choice is the 0-1 loss

$$\ell(h,x,u) = \mathbb{1}_{\{(f(x)-f(u))h(x,u)<0\}}.$$

U-statistics naturally arise in such settings as for example in Clémençon et al. [2008] where Clémençon and al. study the consistency of the empirical risk minimizer of ranking problems using the theory of U-processes in an i.i.d. framework.

Example: `Bipartite ranking problems`
We describe the concrete problem of bipartite ranking. We consider that we have as input a training set of examples. Each example is described by some feature vector and is associated with a binary label. Typically one can consider that we have access to health data of an individual along time. We know at each time step her/his health status x_t and her/his label which is 0 if the individual is healthy and 1 if she/he is sick. In the bipartite ranking problem, we want to learn a `scorer` *which maps any feature vector describing the health status of the individual to a real number such that sick states have a higher score than healthy ones. Following the health status of individuals is time-consuming and we cannot afford to wait for the end of the data acquisition process to understand the relationship between the feature vector describing the health status of the individual and her/his sickness. In such settings where data is coming on the fly, online algorithms are common tools that allow to learn a scorer function along time. At each time step the scorer function is updated based on the new measurement provided.*

5.4.1.3 Generalization bounds for online learning

The performance of online learning algorithms is often analyzed with the notion of *regret* which compares the payoff obtained by the algorithm along time with the one that would have been obtained by taking the optimal decision at each time step [cf. Bubeck and Cesa-Bianchi, 2012, Hoi et al., 2021]. It is natural to wonder if stronger theoretical guarantees can be obtained when some probabilistic structure underlies the sequence of examples, or loss functions, presented to the online algorithm. As asked in Agarwal and Duchi [2012], *"if the sequence of examples are generated by a stochastic process, can the online learning algorithm output a good predictor for future samples from the same process?"* In other words, we want to study the generalization ability of some online learner that generates a sequence of hypothesis $(h_t)_{t\geq 1}$ by bounding with high probability the excess risk defined as

$$\frac{1}{n}\sum_{t=1}^{n} \mathcal{R}(h_t) - \min_{h\in\mathcal{H}} \mathcal{R}(h).$$

Generalization bounds for online learning with pairwise loss functions working with i.i.d. samples have been considered for quite a while in both Machine Learning and Statistics literature [cf. Chen and Lei, 2018, Guo et al., 2017, Kar et al., 2013, Ying and Zhou, 2017]. For dependent data sequences, generalization bounds for online algorithms have also been proved in the last decades with univariate loss functions [cf. Agarwal and Duchi, 2012, Xu et al., 2014, Zhang, 2005]. However, theoretical guarantees for the generalization performance of online algorithms with pairwise loss functions with non i.i.d. data have been so far understudied. A quick and incomplete review of the literature is presented in Table 5.1.

	Univariate loss function	Pairwise loss function
i.i.d. data	Hoi et al., Section 3.7 and references therein	Chen and Lei, Guo et al., Kar et al., Ying and Zhou
Dependent data	Agarwal and Duchi, Xu et al., Zhang	Our work

Table 5.1: Overview of the literature providing generalization bounds for online learning algorithms.

5.4.1.4 Generalization bounds for pairwise online learning with dependent data

Connection with the existing literature. As far as we know, the few papers that investigate the generalization performance of pairwise online learning algorithms with non i.i.d. data have studied specific algorithms and/or specific learning tasks [cf. Qin et al., 2021, Zeng et al., 2021]. In Zeng et al. [2021], the authors analyze online pairwise support vector machine while the work Qin et al. [2021] is focused on online regularized pairwise learning algorithm with least squares loss function. One possible reason explaining this gap in the literature is that *"for pairwise learning [where] pairs of training examples are not i.i.d., [...] standard techniques can not be directly applied."* [cf. Zeng et al., 2021].
With the upcoming application, we are the first - as far as we know - to provide a generalization bounds for online algorithms with pairwise loss functions and Markov chain samples that hold for an arbitrary online learner, covering a large span of settings.

Online learning with a Markovian dynamic. The theoretical analysis of Machine Learning algorithms with an underlying Markovian distribution of the data has become a very active field of research. The first papers to study online learning with samples drawn from non-identical distributions were Smale and Zhou [2009] and Steinwart et al. [2009] where online learning for least square regression and off-line support vector machines are investigated. In Zou et al. [2009], the generalization performance of the empirical risk minimization algorithm is studied with uniformly ergodic Markov chain samples. Hence the analysis of online algorithms with dependent samples is recent and several works make the assumption that the sequence is a uniformly ergodic Markov chain. We motivate the Markovian assumption on the example of the previous paragraph.

Example (continued): `Interest in online algorithms with Markovian dynamic`
The health status of the individual at time $n+1$ is not independent from the past and a simple way to model this time evolution would be to consider that it only depends on the last measured health status namely the feature vector x_n. This is a Markovian assumption on the sequence of observed health status of the individual.

We have explained why pairwise loss functions capture ranking problems and naturally arise in several Machine Learning problems such as metric learning or bipartite ranking (see Clémençon et al. [2008]). We have shown the interest to provide a generalization bounds for online learning with pairwise loss functions with a Markovian assumption on the distribution of the sequence of examples and this is the goal of the next section.

5.4.2 Online-to-batch conversion for pairwise loss functions with Markov chains

We consider a reversible Markov chain $(X_i)_{i\geq 1}$ with state space E satisfying Assumption 1 with stationary distribution π. Using [Meyn and Tweedie, 1993, Theorem 16.0.2], we deduce that there exist constants $0<\rho<1$ and $L>0$ such that

$$\|P^n(x,\cdot)-\pi\|_{\mathrm{TV}} \leq L\rho^n, \qquad \forall n\geq 0,\ \pi\text{−a.e } x\in E. \tag{5.6}$$

We assume that we have a function $f: E\to\mathbb{R}$ which defines the ordering of the objects in E. We aim at finding a relevant approximation of the ordering of the objects in E by selecting a function h (called a *hypothesis* function) in a space $\mathcal{H}$ based on the observation of the random sequence $(X_i, f(X_i))_{1\leq i\leq n}$. To measure the performance of a given hypothesis $h: E\times E\to\mathbb{R}$, we use a pairwise loss function of the form $\ell(h,X,U)$. Typically, one could use the *misranking loss* defined by

$$\ell(h,x,u)=\mathbb{1}_{\{(f(x)-f(u))h(x,u)<0\}},$$

which is 1 if the examples are ranked in the wrong order and 0 otherwise. The goal of the learning problem is to find a hypothesis h which minimizes the *expected misranking risk*

$$\mathcal{R}(h) := \mathbb{E}_{(X,X')\sim\pi\otimes\pi}\Big[\ell(h,X,X')\Big].$$

We show that the investigation of the generalization performance of online algorithms with pairwise loss functions provided by Wang et al. [2012] can be extended to a Markovian framework. Our contribution is two fold.

- Firstly, we prove that with high probability, the average risk of the sequence of hypotheses generated by an arbitrary online learner is bounded by some easily computable statistic.
- This first technical result is then used to show how we can extract a low risk hypothesis from a given sequence of hypotheses selected by an online learner. This is an *online-to-batch* conversion for pairwise loss functions with a Markovian assumption on the distribution of the observed states.

Given a sequence of hypotheses $(h_i)_{1\leq i\leq n} \in \mathcal{H}^n$ generated by any online algorithm, we define the *average paired empirical risk* $\mathcal{M}^n(h_1,\dots,h_{n-1-b_n})$ (see Eq.(5.7)) averaging the *paired empirical risks* M_t (see Eq.(5.8)) of hypotheses h_{t-b_n} when paired with X_t as follows

$$\mathcal{M}^n(h_1,\dots,h_{n-1-b_n}) := \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} M_t, \tag{5.7}$$

$$\text{and} \quad M_t := \frac{1}{t-b_n}\sum_{i=1}^{t-b_n} \ell(h_{t-b_n},X_t,X_i), \tag{5.8}$$

where

$$c_n = \lceil c\times n\rceil \text{ for some } c\in(0,1) \quad \text{and} \quad b_n = \lfloor q\log(n)\rfloor, \tag{5.9}$$

for an arbitrarily chosen $q > \frac{1}{\log(1/\rho)}$ where ρ is a constant related to the uniform ergodicity of the Markov chain, see Eq.(5.6). In the following, we will simply denote $\mathcal{M}^n(h_1,\dots,h_{n-1-b_n})$ by $\mathcal{M}^n$ when the sequence of considered hypotheses is clear from the context.
M_t is the *paired empirical risk* of hypothesis h_{t-b_n} with X_t. It measures the performance of the hypothesis h_{t-b_n} on the example X_t when paired with examples seen before time $t-b_n$. $\mathcal{M}^n$ is the mean value of a proportion $1-c$ of these paired empirical risks. Hence the parameter $c\in(0,1)$ controls the proportion of hypotheses h_{t-b_n} whose paired empirical risk M_t does not appear in the average paired empirical risk value $\mathcal{M}^n$. The parameter b_n controls the time gaps between elements of pairs (X_t,X_i) appearing in Eq.(5.8) in such way that their joint law is close to the product law $\pi\otimes\pi$ (mixing of the chain is met). The use of the burning parameter b_n is the main difference with the work Wang et al. [2012] when defining $\mathcal{M}^n$ and M_t in Eq.(5.7) and Eq.(5.8). From a pragmatic point of view,

- we discard the first hypotheses that are not reliable, namely we do not consider hypothesis h_i for $i \leq c_n - b_n$. These first hypotheses are considered as not reliable since the online learner selected them based on a too small number of observed examples.
- since h_{t-b_n} is learned from $X_1,\dots,X_{t-b_n}$, we test the performance of h_{t-b_n} on X_t (and not on some X_i with $t-b_n+1\leq i<t$) to ensure that the distribution of X_t conditionally on $\sigma(X_1,\dots,X_{t-b_n})$ is approximately the stationary distribution of the chain π (see Assumption 1 and Equation (5.6)). Stated otherwise, this ensures that sufficient mixing has occurred.

Note that we assume n large enough to ensure that $c_n - b_n \geq 1$. For any $\eta > 0$, we denote $\mathcal{N}(\mathcal{H},\eta)$ the L^∞ η-covering number for the hypothesis class $\mathcal{H}$ (see Definition 5.6).

Definition 5.6. [cf. Wainwright, 2019, Chapter 5.1] Let us consider some $\eta > 0$. A L^∞ η-cover of a set $\mathcal{H}$ is a set $\{g_1,\dots,g_N\}\subset\mathcal{H}$ such that for any $h\in\mathcal{H}$, there exists some $i\in\{1,\dots,N\}$ such that $\|g_i-h\|_\infty\leq\eta$. The L^∞ η-covering number $\mathcal{N}(\mathcal{H},\eta)$ is the cardinality of the smallest L^∞ η-cover of the set $\mathcal{H}$.

Theorem 5.7 bounds the average risk of the sequence of hypotheses in terms of its empirical counterpart $\mathcal{M}^n$ and is proved in Section 5.7.1.

Theorem 5.7. *Assume that the Markov chain $(X_i)_{i\geq 1}$ is reversible and satisfies Assumption 1. Assume the hypothesis space $(\mathcal{H}, \|\cdot\|_\infty)$ is compact. Let $h_0, h_1, \ldots, h_n \in \mathcal{H}$ be the ensemble of hypotheses generated by an arbitrary online algorithm working with a pairwise loss function ℓ such that,*

$$\ell(h, x_1, x_2) = \phi(f(x_1) - f(x_2), h(x_1, x_2)),$$

where $\phi : \mathbb{R}\times\mathbb{R} \to [0,1]$ is a Lipschitz function w.r.t. the second variable with a finite Lipschitz constant $Lip(\phi)$. Let $\xi > 0$ be an arbitrary positive number and let us consider $q = \frac{\xi+1}{\log(1/\rho)}$ for the definition of b_n (see Eq.(5.9)). Then for all $c > 0$ and for all $\epsilon > 0$ such that $\epsilon \underset{n\to\infty}{=} o\left(n^{\xi}\right)$, we have for sufficiently large n

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n}) - \mathcal{M}^n\right| \geq \epsilon\right) \\
\leq\quad & 2\left[32\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8Lip(\phi)}\right) + 1\right] b_n \exp\left(-\frac{(c_n - b_n)C(m,\tau)\epsilon^2}{16 b_n^2}\right),
\end{aligned}$$

where $C(m,\tau)^{-1} = 7\times 10^3 \times m^2\tau^2$. We refer to Assumption 1 and the following remark for the definitions of the constants m and τ that depend on the Markov chain $(X_i)_{i\geq 1}$.

Theorem 5.7 shows that average paired empirical risk $\mathcal{M}^n$ (see Eq.(5.7)) is close to average risk given by

$$\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})\,.$$

Quantitative errors bounds can be given assuming that the L^∞-metric entropy (l.h.s of the next equation) satisfies

$$\log\mathcal{N}(\mathcal{H},\eta) = \mathcal{O}(\eta^{-\theta})\,, \tag{5.10}$$

where $\theta > 0$ is an exponent, depending on the dimension of state space E and the regularity of hypotheses of $\mathcal{H}$, that can be computed in some situations (Lipschitz function, higher order smoothness classes, see [Wainwright, 2019, Chapter 5.1] for instance). Theorem 5.9 made this statement rigorous (cf. Eq.(5.11)).
As previously mentioned, online learning algorithms are often studied through the lens of *regret*. The definition of a regret bound in our context is provided in Definition 5.8.

Definition 5.8. An online learning algorithm will be said to have a regret bound $\mathfrak{R}_n$ if it presents an ensemble $h_1, \ldots, h_{n-1}$ such that

$$\mathcal{M}^n \leq \min_{h\in\mathcal{H}}\left\{\mathcal{M}^n(h,\ldots,h)\right\} + \mathfrak{R}_n.$$

In the literature of learning theory Cucker and Zhou [2007], we are often interested in the averaged excess generalization error

$$\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n}) - \mathcal{R}(h^*),$$

where h^* is the population risk minimizer and is given by $h^* \in \underset{h\in\mathcal{H}}{\arg\min}\ \mathcal{R}(h)$. As a consequence, most of works focused on online-to-batch conversion are interested in the overall convergence rate of the excess generalization error for online learners that achieve a given regret bound. Examples can be found with [Guo et al., 2017, Corollary 4] or with [Kar et al., 2013, Theorem 5] where both papers work with pairwise loss functions with i.i.d. observations. In Theorem 5.9 (cf. Eq.(5.12)) we provide the overall rate for the averaged excess generalization error for an online learning satisfying a given regret bound. Theorem 5.9 is proved in Section 5.7.2 and should be understood as an extension of the above mentioned results from Kar et al. [2013] and Guo et al. [2017].

Theorem 5.9. *We keep the notations and assumptions of Theorem 5.7. Assume further that $\mathcal{H}$ satisfies Eq.(5.10). Then it holds*

$$\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n}) - \mathcal{M}^n\right| = \mathcal{O}_{\mathbb{P}}\left[\frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}}\right]. \tag{5.11}$$

Moreover, if the online learner has a regret bound $\mathfrak{R}_n$ (cf. Definition 5.8), it holds

$$\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\mathcal{R}(h^*)=\mathcal{O}_{\mathbb{P}}\left[\frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}}+\mathfrak{R}_n\right]. \tag{5.12}$$

5.4.3 Batch hypothesis selection

Theorems 5.7 and 5.9 are results on the performance of online learning algorithms. We will use these results to study the generalization performance of such online algorithms in the batch setting (see Theorem 5.10). Hence we are now interested in *selecting a good hypothesis from the ensemble of hypotheses generated by the online learner* namely that has a small empirical risk.
We measure the risk for h_{t-b_n} on the last $n-t$ examples of the sequence $X_1,\dots,X_n$, and penalize each h_{t-b_n} based on the number of examples on which it is evaluated. More precisely, let us define the empirical risk of hypothesis h_{t-b_n} on $\{X_{t+1},\dots,X_n\}$ as

$$\widehat{\mathcal{R}}(h_{t-b_n},t+1):=\binom{n-t}{2}^{-1}\sum_{k>i,i\geq t+1}^{n}\ell(h_{t-b_n},X_i,X_k).$$

For a confidence parameter $\gamma\in(0,1)$ that will be specified in Theorem 5.10, the hypothesis $\widehat{h}$ is chosen to minimize the following penalized empirical risk,

$$\widehat{h}=h_{\widehat{t}-b_n}\quad\text{and}\quad\widehat{t}\in\arg\min_{c_n\leq t\leq n-1}\left(\widehat{\mathcal{R}}(h_{t-b_n},t+1)+c_\gamma(n-t)\right), \tag{5.13}$$

where

$$c_\gamma(x)=\sqrt{\frac{C(m,\tau)^{-1}}{x}\log\frac{64(n-c_n)(n-c_n+1)}{\gamma}},$$

with $C(m,\tau)^{-1}=7\times10^3\times m^2\tau^2$.
Theorem 5.10 proves that the model selection mechanism previously described select a hypothesis $\widehat{h}$ from the hypotheses of an arbitrary online learner whose risk is bounded relative to $\mathcal{M}^n$. The proof of Theorem 5.10 is postponed to Section 5.7.3.

Theorem 5.10. *Assume that the Markov chain $(X_i)_{i\geq1}$ is reversible and satisfies Assumptions 1 and 2. Let $h_0,\dots,h_n$ be the set of hypotheses generated by an arbitrary online algorithm $\mathcal{A}$ working with a pairwise loss ℓ which satisfies the conditions given in Theorem 5.7. Let $\xi>0$ be an arbitrary positive number and let us consider $q=\frac{\xi+1}{\log(1/\rho)}$ for the definition of b_n (see Eq.(5.9)). For all $\epsilon>0$ such that $\epsilon\underset{n\to\infty}{=}o\left(n^{\xi}\right)$, if the hypothesis is selected via Eq.(5.13) with the confidence γ chosen as*

$$\gamma=64(n-c_n+1)\exp\left(-(n-c_n)\epsilon^2C(m,\tau)/128\right),$$

then, when n is sufficiently large, we have

$$\mathbb{P}\left(\mathcal{R}(\widehat{h})\geq\mathcal{M}^n+\epsilon\right)\leq32\left[\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{16\mathrm{Lip}(\phi)}\right)+1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{(16b_n)^2}+2\log n\right).$$

Analogously to the previous section, we can derive from Theorem 5.10 a bound for the excess risk of the selected hypothesis $\widehat{h}$.

Corollary 5.11. *We keep the notations and assumptions of Theorem 5.10. Assume further that $\mathcal{H}$ satisfies Eq.(5.10). Then it holds*

$$\mathcal{R}(\widehat{h})-\mathcal{M}^n=\mathcal{O}_{\mathbb{P}}\left[\frac{\log^2 n}{n^{\frac{1}{2+\theta}}}\right].$$

Moreover, if the online learner has a regret bound $\mathfrak{R}_n$ (cf. Definition 5.8), it holds

$$\mathcal{R}(\widehat{h})-\mathcal{R}(h^*)=\mathcal{O}_{\mathbb{P}}\left[\frac{\log^2 n}{n^{\frac{1}{2+\theta}}}+\mathfrak{R}_n\right].$$

5.5 Adaptive goodness-of-fit tests in a density model

5.5.1 Goodness-of-fit tests and review of the literature

In its original formulation, the goodness-of-fit test aims at determining if a given distribution q matches some unknown distribution p from samples $(X_i)_{i\geq 1}$ drawn independently from p. Classical approaches to solve the goodness-of-fit problem use the empirical process theory. Most of the popular tests such as the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics are based on the empirical distribution function of the samples. Other traditional approaches may require space partitioning or closed-form integrals Baringhaus and Henze [1988], Beirlant et al. [2008]. In Rudzkis and Bakshaev [2013], a non-parametric method is proposed with a test based on a kernel density estimator. In the last decade, a lot of effort has been put into finding more efficient goodness-of-fit tests. The motivation was mainly coming from graphical models where the distributions are known up to a normalization factor that is often computationally intractable. To address this problem, several tests have been proposed based on Reproducing Kernel Hilbert Space (RKHS) embedding. A large span of them use classes of Stein transformed RKHS functions [Gorham and Mackey, 2017, Liu et al., 2016]. For example in Chwialkowski et al. [2016], a goodness-of-fit test is proposed for both i.i.d or non i.i.d samples. The test statistic uses the squared Stein discrepancy, which is naturally estimated by a V-statistic. One drawback of such approach is that the theoretical results provided are only asymptotic. This paper is part of a large list of works that proposed a goodness-of-fit test and where the use of U-statistics naturally emerge [cf. Butucea et al., 2007, Fan, 1997, Fan and Ullah, 1999, Fernández and Gretton, 2019, Fromont and Laurent, 2006, Liu et al., 2016]. To conduct a non-asymptotic analysis of the goodness-of-fit tests proposed for non i.i.d samples, a concentration result for U-statistics with dependent random variables is much needed.

5.5.2 Goodness-of-fit test for the density of the stationary measure of a Markov chain

In this section, we provide a goodness-of-fit test for Markov chains whose stationary distribution has density with respect to the Lebesgue measure λ_{Leb} on $\mathbb{R}$. Our work is inspired from Fromont and Laurent [2006] where Fromont and Laurent tackled the goodness-of-fit test with i.i.d samples. Conducting a non-asymptotic theoretical study of our test, we are able to identify the classes of alternatives over which our method has a prescribed power.
Let $X_1, \dots, X_n$ be a Markov chain with stationary distribution π with density f with respect to the Lebesgue measure on $\mathbb{R}$. Let f_0 be some given density in $L^2(\mathbb{R})$ and let α be in $]0,1[$. Assuming that f belongs to $L^2(\mathbb{R})$, we construct a level α test of the null hypothesis "$f = f_0$" against the alternative "$f \neq f_0$" from the observation $(X_1, \dots, X_n)$. The test is based on the estimation of $\|f - f_0\|_2^2$ that is $\|f\|_2^2 + \|f_0\|_2^2 - 2\langle f, f_0\rangle$. $\langle f, f_0\rangle$ is usually estimated by the empirical estimator $\sum_{i=1}^n f_0(X_i)/n$ and the cornerstone of our approach is to find a way to estimate $\|f\|_2^2$. We follow the work of Fromont and Laurent [2006] and we introduce a set $\{S_m, m \in \mathcal{M}\}$ of linear subspaces of $L^2(\mathbb{R})$. For all m in $\mathcal{M}$, let $\{p_l, l \in \mathcal{L}_m\}$ be some orthonormal basis of S_m. The variable

$$\widehat{\theta}_m = \frac{1}{n(n-1)} \sum_{l \in \mathcal{L}_m} \sum_{i \neq j = 1}^{n} p_l(X_i) p_l(X_j)$$

estimates $\|\Pi_{S_m}(f)\|_2^2$ where Π_{S_m} denotes the orthogonal projection onto S_m. Then $\|f - f_0\|_2^2$ can be approximated by

$$\widehat{T}_m = \widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n} \sum_{i=1}^{n} f_0(X_i),$$

for any m in $\mathcal{M}$. Denoting by $t_m(u)$ the $(1-u)$ quantile of the law of $\widehat{T}_m$ under the hypothesis "$f = f_0$" and considering

$$u_\alpha = \sup_{u \in]0,1[} \mathbb{P}_{f_0}\left(\sup_{m \in \mathcal{M}} (\widehat{T}_m - t_m(u)) > 0 \right) \leq \alpha,$$

we introduce the test statistic T_α defined by

$$T_\alpha = \sup_{m\in\mathcal{M}} (\widehat{T}_m - t_m(u_\alpha)). \tag{5.14}$$

The test consists in rejecting the null hypothesis if T_α is positive. This approach can be read as a multiple testing procedure. Indeed, for each m in $\mathcal{M}$, we construct a level u_α test of the null hypothesis "$f = f_0$" by rejecting this hypothesis if $\widehat{T}_m$ is larger than its $(1-u_\alpha)$ quantile under the hypothesis "$f = f_0$". We thus obtain a collection of tests and we decide to reject the null hypothesis if for some of the tests of the collection this hypothesis is rejected.
Now we define the different collection of linear subspaces $\{S_m, m \in \mathcal{M}\}$ that we will use in the following. We will focus on constant piecewise functions, scaling functions and, in the case of compactly supported densities, trigonometric polynomials.

- For all D in $\mathbb{N}^*$ and $k \in \mathbb{Z}$, let
$$I_{D,k} = \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}.$$
For all $D \in \mathbb{N}^*$, we define $S_{(1,D)}$ as the space generated by the functions $\{I_{D,k}, k \in \mathbb{Z}\}$ and
$$\widehat{\theta}_{(1,D)} = \frac{1}{n(n-1)} \sum_{k\in\mathbb{Z}} \sum_{i\neq j=1}^{n} I_{D,k}(X_i) I_{D,k}(X_j).$$

- Let us consider a pair of compactly supported orthonormal wavelets (ϕ, ψ) such that for all $J \in \mathbb{N}$, $\{\phi_{J,k} = 2^{J/2}\phi(2^J \cdot -k), k \in \mathbb{Z}\} \cup \{\psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k), j \in \mathbb{N}, j \geq J, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$. For all $J \in \mathbb{N}$ and $D = 2^J$, we define $S_{(2,D)}$ as the space generated by the scaling functions $\{\phi_{J,k}, k \in \mathbb{Z}\}$ and
$$\widehat{\theta}_{(2,D)} = \frac{1}{n(n-1)} \sum_{k\in\mathbb{Z}} \sum_{i\neq j=1}^{n} \phi_{J,k}(X_i) \phi_{J,k}(X_j).$$

- Let us consider the Fourier basis of $L^2([0,1])$ given by
$$\begin{aligned} g_0(x) &= \mathbb{1}_{[0,1]}(x), \\ g_{2p-1}(x) &= \sqrt{2}\cos(2\pi p x)\mathbb{1}_{[0,1]}(x) \quad \forall p \geq 1, \\ g_{2p}(x) &= \sqrt{2}\sin(2\pi p x)\mathbb{1}_{[0,1]}(x) \quad \forall p \geq 1. \end{aligned}$$
For all $D \in \mathbb{N}^*$, we define $S_{(3,D)}$ as the space generated by the functions $\{g_l, l = 0, \ldots, D\}$ and
$$\widehat{\theta}_{(3,D)} = \frac{1}{n(n-1)} \sum_{l=0}^{D} \sum_{i\neq j=1}^{n} g_l(X_i) g_l(X_j).$$

We denote $\mathbb{D}_1 = \mathbb{D}_3 = \mathbb{N}^*$ and $\mathbb{D}_2 = \{2^J, J \in \mathbb{N}\}$. For l in $\{1,2,3\}$, D in $\mathbb{D}_l$, $\Pi_{S_{(l,D)}}$ denotes the orthogonal projection onto $S_{(l,D)}$ in $L^2(\mathbb{R})$. For all l in $\{1,2,3\}$, we take $\mathcal{D}_l \subset \mathbb{D}_l$ with $\cup_{l\in\{1,2,3\}} \mathcal{D}_l \neq \emptyset$ and $\mathcal{D}_3 = \emptyset$ if the X_i's are not included in $[0,1]$.
Let $\mathcal{M} = \{(l,D), l \in \{1,2,3\}, D \in \mathcal{D}_l\}$. Theorem 5.12 describes classes of alternatives over which the corresponding test has a prescribed power. We work under the additional Assumption 4.(ii). We refer to Section 5.8.1 for the proof of Theorem 5.12.

Assumption 4.(ii) (cf. Section 4.2.5 in Chapter 4)
The initial distribution of the Markov chain $(X_i)_{i\geq 1}$, denoted χ, is absolutely continuous with respect to the stationary measure π and its density, denoted by $\frac{d\chi}{d\pi}$, has finite p-moment for some $p \in (1,\infty]$, i.e

$$\infty > \left\| \frac{d\chi}{d\pi} \right\|_{\pi,p} := \begin{cases} \left[\int \left| \frac{d\chi}{d\pi} \right|^p d\pi \right]^{1/p} & \text{if } p < \infty, \\ \operatorname{ess\,sup} \left| \frac{d\chi}{d\pi} \right| & \text{if } p = \infty. \end{cases}$$

In the following, we will denote $q = \frac{p}{p-1} \in [1,\infty)$ (with $q = 1$ if $p = +\infty$) which satisfies $\frac{1}{p} + \frac{1}{q} = 1$.

Theorem 5.12. *Let $X_1, \dots, X_n$ a Markov chain on $\mathbb{R}$ satisfying the Assumptions 1, 2 and 4.(ii) with stationary measure π. We assume that π has density f with respect the Lebesgue measure on $\mathbb{R}$ and let f_0 be some given density. Let T_α be the test statistic defined by Eq.(5.14). Assume that f_0 and f belong to $L^\infty(\mathbb{R})$ (the space of essentially bounded measurable functions on $\mathbb{R}$) and that there exist $p_1, p_2 \in (1, +\infty]$ such that*

$$C_\chi := \left\| \frac{1}{f} \frac{d\chi}{d\lambda_{Leb}} \right\|_{f\lambda_{Leb}, p_1} \vee \left\| \frac{1}{f_0} \frac{d\chi}{d\lambda_{Leb}} \right\|_{f_0\lambda_{Leb}, p_2} < \infty,$$

where we used the notations of Assumption 4.(ii). We fix some γ in $]0,1[$. For any $\epsilon \in]0,2[$, there exist some positive constants C_1, C_2, C_3 such that, setting for all $m = (l, D)$ in $\mathcal{M}$,

$$V_m(\gamma) = C_1 \|f\|_\infty \frac{\log(3C_\chi/\gamma)}{\epsilon n} + C_2 \left(\|f\|_\infty \log(D+1) + \|f_0\|_\infty \right) \frac{\log(3C_\chi/\gamma)}{n} + C_3 \left(\|f\|_\infty + 1 \right) DR\left(n, \log\left\{ \frac{3\beta \log n}{\gamma} \right\} \right),$$

with

$$R(n,u) = \log n \left\{ \frac{u}{n} + \left(\frac{u}{n} \right)^2 \right\},$$

if f satisfies

$$\|f - f_0\|_2^2 > (1+\epsilon) \inf_{m \in \mathcal{M}} \left\{ \|f - \Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha) + V_m(\gamma) \right\}, \tag{5.15}$$

then

$$\mathbb{P}_f \left(T_\alpha \leq 0 \right) \leq \gamma.$$

In order to make the condition (5.15) more explicit and to study its sharpness, we define the uniform separation rate which provides for any $\gamma \in (0,1)$ the smallest distance between the set of null hypotheses and the set of alternatives to ensure that the power of our statistic test with level α is at least $1-\gamma$.

Definition 5.13. Given $\gamma \in]0,1[$ and a class of functions $\mathcal{B} \subset L^2(\mathbb{R})$, we define the uniform separation rate $\rho(\Phi_\alpha, \mathcal{B}, \gamma)$ of a level α test Φ_α of the null hypothesis "$f \in \mathcal{F}$" over the class $\mathcal{B}$ as the smallest number ρ such that the test guarantees a power at least equal to $(1-\gamma)$ for all alternatives $f \in \mathcal{B}$ at a distance ρ from $\mathcal{F}$. Stated otherwise, denoting by $d_2(f, \mathcal{F})$ the L^2-distance between f and $\mathcal{F}$ and by $\mathbb{P}_f$ the distribution of the observation $(X_1, \dots, X_n)$,

$$\rho(\Phi_\alpha, \mathcal{B}, \gamma) = \inf \left\{ \rho > 0, \forall f \in \mathcal{B}, d_2(f, \mathcal{F}) \geq \rho \implies \mathbb{P}_f(\Phi_\alpha \text{ rejects}) \geq 1 - \gamma \right\}.$$

In the following, we derive on explicit upper bound on the uniform separation rates of the test proposed above over several classes of alternatives. For $s > 0, P > 0, M > 0$ and $l \in \{1,2,3\}$, we introduce

$$\mathcal{B}_s^{(l)}(P, M) = \left\{ f \in L^2(\mathbb{R}) \mid \forall D \in \mathcal{D}_l, \quad \|f - \Pi_{S_{(l,D)}}(f)\|_2^2 \leq P^2 D^{-2s}, \|f\|_\infty \leq M \right\}.$$

These sets of functions include some Hölder balls or Besov bodies with smoothness s, as highlighted in Fromont and Laurent [2006, Section 2.3]. Corollary 5.14 gives an upper bound for the uniform separation rate of our testing procedure over the classes $\mathcal{B}_s^{(l)}(P, M)$ and is proved in Section 5.8.3.

Corollary 5.14. *Let T_α be the test statistic defined by (5.14). Assume that for $l \in \{1,2,3\}$, $\mathcal{D}_l$ is $\{2^J, 0 \leq J \leq \log_2\left(n/(\log(n)\log\log n)^2\right)\}$ or $\emptyset$. For all $s > 0$, $M > 0, P > 0$ and $l \in \{1,2,3\}$ such that $\mathcal{D}_l \neq \emptyset$, there exists some positive constant $C = C(s, \alpha, \gamma, M, \|f_0\|_\infty)$ such that the uniform separation rate of the test $\mathbb{1}_{T_\alpha > 0}$ over $\mathcal{B}_s^{(l)}(P, M)$ satisfies for n large enough*

$$\rho\left(\mathbb{1}_{T_\alpha > 0}, \mathcal{B}_s^{(l)}(P, M), \gamma \right) \leq C' P^{\frac{1}{2s+1}} \left(\frac{\log(n)\log\log n}{n} \right)^{\frac{s}{2s+1}}.$$

Remark. In Corollary 5.14, the condition *n large enough* corresponds to

$$\left(\log(n) \frac{\log\log n}{n} \right)^{1/2} \leq P \leq \frac{n^s}{(\log(n)\log\log n)^{2s+1/2}}.$$

For the problem of testing the null hypothesis "$f = \mathbb{1}_{[0,1]}$" against the alternative $f = \mathbb{1}_{[0,1]} + g$ with $g \neq 0$ and $g \in B_s(P)$ where $B_s(P)$ is a class of smooth functions (like some Hölder, Sobolev

or Besov ball in $L^2([0,1])$) with unknown smoothness parameter s, Ingster [1993] established in the case where the random variables $(X_i)_{i\geq 1}$ are i.i.d. that the adaptive minimax rate of testing is of order $(\sqrt{\log\log n}/n)^{2s/(4s+1)}$. From Corollary 5.14, we see that our procedure leads to a rate which is close (at least for sufficiently large smoothness parameter s) to the one derived by Ingster in the i.i.d. framework since the upper bound on the uniform separation rate from Corollary 5.14 can be read (up to a log factor) as $([\log\log n]/n)^{\frac{2s}{4s+2}}$.

5.5.3 Simulations

We propose to test our method on three practical examples. The code is available at `https://github.com/quentin-duchemin/goodness-of-fit-MC`. In all our simulations, we use Markov chains of length $n = 100$. We choose different alternatives to test our method and we use i.i.d. samples from these distributions. We chose a level $\alpha = 5\%$ for all our experiments. All tests are conducted as follows.

1. We start by the estimation of the $(1-u)$ quantiles $t_m(u)$ of the variables $\widehat{T}_m = \widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^n f_0(X_i)$ under the hypothesis "$f = f_0$" for u varying on a regular grid of $]0,\alpha[$. We sample $5,000$ sequences of length $n = 100$ with i.i.d. random variables with distribution f_0. We end up with an estimation $\widehat{t}_m(u)$ of $t_m(u)$ for any u in the grid and any $m \in \mathcal{M}$.
2. Then, we estimate the value of u_α. We sample again $5,000$ sequences of length $n = 100$ with i.i.d. random variables with distribution f_0. We use them to estimate the probabilities $\mathbb{P}_{f_0}(\sup_{m\in\mathcal{M}}(\widehat{T}_m - \widehat{t}_m(u)) > 0)$ for any u in the grid and we keep the larger value of u such that the corresponding probability is still larger than α. The selected value of the grid is called u_α. Thanks to the first step, we have the estimates $\widehat{t}_m(u_\alpha)$ of $t_m(u_\alpha)$ for any $m \in \mathcal{M}$.
3. Finally, we sample $5,000$ Markov chains with length $n = 100$ with stationary distribution f. For each sequence, we can compute $\widehat{T}_m$. Dividing by $5,000$ the number of sequences for which $\sup_{m\in\mathcal{M}}(\widehat{T}_m - \widehat{t}_m(u_\alpha)) > 0$, we get an estimation of the power of the test.

To define comparison points, we compare the power of our test with the classical Kolmogorov-Smirnov test (KS test) and the Chi-squared test (χ^2 test). The rejection region associated with a test of level 5% is set by *sampling under the null* for both the KS test and the χ^2 test. With Figure 5.3, we provide a visualization of the density of the stationary distribution of the Markov chain and of the density of the alternative that gives the smaller power on our experiments.

5.5.3.1 Example 1: AR(1) process

Let us consider some $\theta \in (0,1)$. Then, we define the AR(1) process $(X_i)_{i\geq 1}$ starting from $X_1 = 0$ with for any $n \geq 1$,

$$X_{n+1} = \theta X_n + \xi_{n+1},$$

where $(\xi_n)_n$ are i.i.d. random variables with distribution $\mathcal{N}(0,\tau^2)$ with $\tau > 0$. From Example 1 from Section 4.3.3.1, we know that Assumptions 1 and 2 hold. The stationary measure π of the Markov chain $(X_i)_{i\geq 1}$ is $\mathcal{N}\left(0, \frac{\tau^2}{1-\theta^2}\right)$, i.e. π has density f with respect to the Lebesgue measure on $\mathbb{R}$ with

$$\forall y \in \mathbb{R}, \quad f(y) = \frac{\sqrt{1-\theta^2}}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(1-\theta^2)y^2}{2\tau^2}\right).$$

We focus on the following alternatives

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Table 5.2 shows the estimated powers for our test, the KS test and the χ^2 test.

5.5.3.2 Example 2: Markov chain generated from independent Metropolis Hasting algorithm

Let us consider the probability measure π with density f with respect to the Lebesgue measure on $[-3,3]$ where

$$\forall x \in [-3,3], \quad f(x) = \frac{1}{Z} e^{-x^2} \left(3 + \sin(5x) + \sin(2x)\right),$$

(μ, σ^2)	**Our test**	χ^2 test	**KS** test	$\|\mathbf{f} - \mathbf{f}_{\mu,\sigma^2}\|_2$
$(2, 1.5)$	0.99	0.85	0.98	0.39
$(0, 1)$	0.97	0.9	0.8	0.2
$(-0.2, 1.2)$	0.86	0.63	0.84	0.17
$(0, 1.2)$	0.81	0.64	0.82	0.16
$(0, 2)$	0.1	0.03	0.29	0.06

Table 5.2: Estimated powers of the tests for Markov chains with size $n = 100$. We worked with $\tau = 1$, $\theta = 0.8$ and $\mathcal{M} = \{(1, i) : i \in \{1, \dots, 10\}\}$. Hence, the stationary distribution of the chain is approximately $\mathcal{N}(0, 2.8)$. For the χ^2 test, we work on the interval $[-5, 5]$ that we split into 20 regular parts.

with Z a normalization constant such that $\int_{-3}^{3} f(x)dx = 1$. To construct a Markov chain with stationary measure π, we use an independent Metropolis-Hasting algorithm with proposal density $q(x) \propto \exp(-x^2/6)$. Using Proposition 5.4, we get that the above built Markov chain $(X_i)_{i\geq 1}$ satisfies Assumptions 1 and 2. We focus on the following alternatives

$$g_{\mu,\sigma^2}(x) = \frac{1}{Z(\mu, \sigma^2)} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathbb{1}_{[-3,3]}(x),$$

where $Z(\mu, \sigma^2)$ is a normalization constant such that $\int g_{\mu,\sigma^2}(x)dx = 1$. Table 5.3 shows the estimated powers for our test, the KS test and the χ^2 test.

(μ, σ^2)	**Our test**	χ^2 test	**KS** test	$\|\mathbf{f} - \mathbf{g}_{\mu,\sigma^2}\|_2$
$(0, 1)$	0.96	0.91	0.9	0.29
$(0, 0.7^2)$	0.95	0.84	0.93	0.23
$(0.3, 0.7^2)$	0.92	0.87	0.93	0.19

Table 5.3: Estimated powers of the tests for Markov chains with size $n = 100$. We used $\mathcal{M} = \{(1, i) : i \in \{1, \dots, 10\}\}$. For the χ^2 test, we work on the interval $[-3, 3]$ that we split into 20 regular parts.

5.5.3.3 Example 3: ARCH process

Let us consider some $\theta \in (-1, 1)$. We are interested in the simple threshold auto-regressive model $(X_n)_{n\geq 1}$ defined by $X_1 = 0$ and for any $n \geq 1$,

$$X_{n+1} = \theta|X_n| + (1 - \theta^2)^{1/2}\xi_{n+1},$$

where the random variables $(\xi_n)_{n\geq 2}$ are i.i.d. with standard Gaussian distribution. From Example 3 from Section 4.3.3.1, we know that Assumptions 1 and 2 hold. The transition kernel of the Markov chain $(X_i)_{i\geq 1}$ is

$$\forall x, y \in \mathbb{R}, \quad P(x, dy) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \theta|x|)^2}{2(1-\theta^2)}\right) dy.$$

The stationary distribution π of the Markov chain has density f with respect to the Lebesgue measure on $\mathbb{R}$ with

$$\forall y \in \mathbb{R}, \quad f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \Phi\left(\frac{\theta y}{(1-\theta^2)^{1/2}}\right),$$

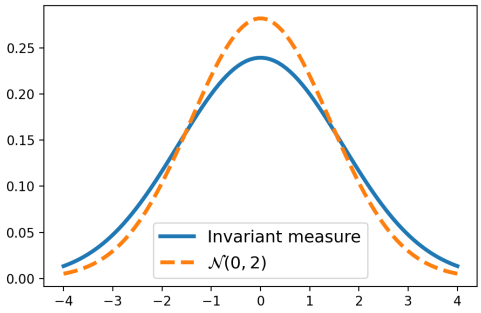
where Φ is the standard normal cumulative distribution function. We focus on the following alternatives

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

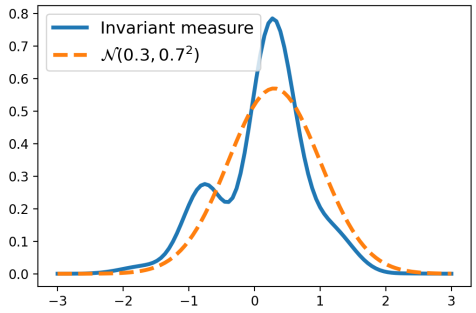
Table 5.4 shows the estimated powers for our test, the KS test and the χ^2 test.

(μ, σ^2)	**Our test**	χ^2 test	**KS** test	$\|\mathbf{f} - \mathbf{f}_{\mu,\sigma^2}\|_2$
$(0, 1)$	0.98	0.85	0.95	0.3
$(1, 0.8^2)$	0.95	0.79	0.88	0.22
$(0.5, 1)$	0.3	0.07	0.5	0.14
$(0.6, 0.8^2)$	0.35	0.16	0.4	0.036

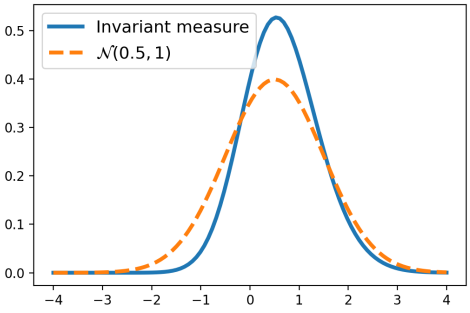
Table 5.4: Estimated powers of the tests for Markov chains with size $n = 100$. We used $\theta = 0.8$ and $\mathcal{M} = \{(1, i) \ : \ i \in \{1, \dots, 10\}\}$. For the χ^2 test, we work on the interval $[-20, 20]$ that we split into 20 regular parts.



(a) Example 1



(b) Example 2



(c) Example 3

Figure 5.3: In solid line, we plot the density of the stationary measure of the Markov chain for the three examples of our simulations. In dotted line, we plot the density of the alternative that gives the smaller power on our experiments.

5.5.3.4 Comments on our numerical experiments

Our experiments show that the χ^2 goodness-of-fit test give in general the smaller power compared to our method and to the KS test. The χ^2 test is better suited to deal with discrete probability distributions and it seems to suffer to small power in our continuous setting. Note that using the χ^2 test with continuous densities on $\mathbb{R}$ require to specify some hyperparameters (such as a compact interval and the number of bins to discretize it). In practice, the test results can change drastically by modifying these hyperparameters, making the test unreliable. Our experiments also show that when the L^2 norm between the true density f and the alternative one f_0 is large enough, our method reaches higher power compared to the two other procedures considered. Nevertheless, our approach seems less powerful compared to the KS test when the L^2 norm $\|f - f_0\|_2$ is getting smaller. This is not surprising since our testing procedure is based on the L^2 norm while the KS test relies on the sup norm between cumulative distribution functions (CDFs). We conduct an final experiment to better stress this distinction between our procedure and the KS test. We consider the notations and the framework of the example

from Section 5.5.3.1 with the following alternatives

$$f^{(L,\delta)}(x) = \begin{cases} f_{0,\frac{\tau^2}{1-\theta^2}}(x) & \text{if } |x| \geq \delta \\ f_{0,\frac{\tau^2}{1-\theta^2}}(x) - L & \text{if } -\delta < x \leq 0 \\ f_{0,\frac{\tau^2}{1-\theta^2}}(x) + L & \text{if } 0 < x < \delta \end{cases},$$

where $L, \delta > 0$ are chosen so that $f^{(L,\delta)}(x) \geq 0$ for any $x \in \mathbb{R}$. We work with $\tau = 1, \theta = 0.8$ and $\mathcal{M} = \{(1,i) \,:\, i \in \{1,\dots,10\}\}$. Figure 5.4 shows the alternatives considered. The sup norm between the CDFs of f and $f^{(L,\delta)}$ is equal to $L\delta$ while the squared L^2 norm between f and $f^{(L,\delta)}$ is $2L^2\delta^3/3$. Hence, we expect that powers will increase for both tests when L and/or δ are increasing. Moreover, we expect the power of our method to be more sensitive to the parameters L and δ. Those intuitions are confirmed with the numerical experiments presented in Table 5.5.

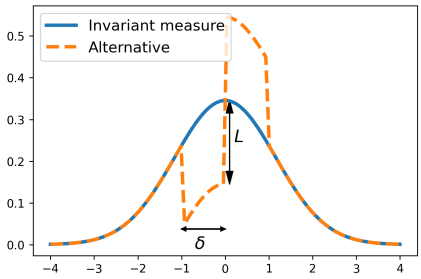


Figure 5.4: Alternative considered.

$(\mathbf{L}, \delta)$	0.25	0.5	0.75	1
0.05	0.06	0.12	0.2	0.21
	0.1	0.15	0.2	0.22
0.05	0.16	0.33	0.36	0.4
	0.23	0.26	0.33	0.37
0.1	0.33	0.66	0.8	0.83
	0.26	0.35	0.46	0.48
0.15	0.82	0.87	0.9	0.95
	0.35	0.45	0.55	0.67
0.2	0.9	0.93	0.95	0.98
	0.46	0.54	0.72	0.87

Table 5.5: Estimated powers of the tests for Markov chains with size $n = 100$. Gray cells are the powers of our method while blank cells are the ones obtained with the KS test.

5.6 Proofs for Section 5.3

Let us explain in a nutshell the structure of our proof. For any natural number R, we denote $\mathbf{H}^R$ the integral operator with kernel function h_R at resolution R, namely

$$h_R(x,y) := \sum_{r \in I, r \leq R} \lambda_r \phi_r(x)\phi_r(y), \quad \mathbf{H}^R f(x) := \int_E h_R(x,y) f(y) d\pi(y).$$

We define $\widetilde{\mathbf{H}}_n^R$ and $\mathbf{H}_n^R$ analogously by using the kernel h_R in Eq.(5.3). Using the triangle inequality, we split the distance $\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))$ into four terms.

1. $\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}^R))$ is a bias term induced by working at resolution R.

2. A non-trivial preliminary work allows to prove that $\delta_2(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R))$ can be written as an empirical process of the Markov chain $(X_i)_{i \geq 1}$ whose tail can be controlled by applying concentration inequalities for sums of functions of uniformly ergodic Markov chains (this is where we use the assumption that Υ is finite). We refer to Eq.(5.16).

3. Since the matrices $\mathbf{H}_n^R$ and $\widetilde{\mathbf{H}}_n^R$ only differ at diagonal elements, $\delta_2(\lambda(\widetilde{\mathbf{H}}_n^R), \lambda(\mathbf{H}_n^R))$ can be coarsely bounded by $n^{-1/2}\|h_R\|_\infty$ (cf. Eq.(5.17)).

4. Applying the Hoffman-Wielandt inequality, one can notice that $\delta_2(\lambda(\mathbf{H}_n^R), \lambda(\mathbf{H}_n))$ can be upper-bounded by a U-statistic of order two of the Markov chain $(X_i)_{i \geq 1}$ (cf. Eq.(5.18)). We control the tail behaviour of this U-statistic by applying Theorem 5.1.

The proof is then concluded by choosing the resolution level R so that $R^2 = \lceil \sqrt{n} \rceil$.

5.6.1 Deviation inequality for the spectrum of signed integral operators

As shown in Section 5.6.2, Theorem 5.3 is a direct consequence of the concentration result provided by Theorem 5.15.

Theorem 5.15. *We keep notations of Section 5.3. Assume that $(X_n)_{n\geq 1}$ is a Markov chain on E satisfying Assumptions 1 and 2 described in Section 5.2.2 with stationary distribution π. Let us consider some symmetric kernel $h : E \times E \to \mathbb{R}$, square integrable with respect to $\pi \otimes \pi$. Let us consider some $R \in \mathbb{N}^*$. We assume that there exist continuous functions $\phi_r : E \to \mathbb{R}$, $r \in I$ (where $I = \mathbb{N}$ or $I = 1, \dots, N$) that form an orthonormal basis of $L^2(\pi)$ such that it holds pointwise*

$$h(x,y) = \sum_{r\in I} \lambda_r \phi_r(x)\phi_r(y),$$

with

$$\Lambda_R := \sup_{r\in I,\, r\leq R} |\lambda_r| \text{ and } \Upsilon_R := \sup_{r\in I,\, r\leq R} \|\phi_r\|_\infty^2.$$

We also define $h_R(x,y) = \sum_{r\in I,\, r\leq R} \lambda_r\phi_r(x)\phi_r(y)$ and we assume that $\|h_R\|_\infty, \|h - h_R\|_\infty < \infty$. Then there exists a universal constant $K > 0$ such that for any $t > 0$, it holds

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2 \geq \left(\|h_R\|_\infty^2 + \kappa\|h-h_R\|_\infty^2\right)\frac{\log n}{n} + 2\sum_{i>R, i\in I}\lambda_i^2 + t\right)\\
\leq\ & 16\exp\left(-n\frac{t^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right) + \beta\log(n)\exp\left(-\frac{n}{16\log n}\left\{\left[\frac{t}{c}\right]\wedge\left[\frac{t}{c}\right]^{1/2}\right\}\right)\\
+\ & 16R^2\exp\left(-\frac{nt}{Km^2\tau^2R^2\Lambda_R^2\Upsilon_R^2}\right).
\end{aligned}$$

where $c = \kappa\|h - h_R\|_\infty$ with $\kappa > 0$ depending on δ_M, τ, L, m and ρ. β depends only on ρ.

Proof of Theorem 5.15. For any integer $R \geq 1$, we denote

$$\begin{aligned}
X_{n,R} &:= \frac{1}{\sqrt{n}}\left(\phi_r(X_i)\right)_{1\leq i\leq n,\ 1\leq r\leq R} \in \mathbb{R}^{n\times R}\\
A_{n,R} &:= \left(X_{n,R}^\top X_{n,R}\right)^{1/2} \in \mathbb{R}^{R\times R}\\
K_R &:= \mathrm{Diag}(\lambda_1,\dots,\lambda_R)\\
\widetilde{\mathbf{H}}_n^R &:= X_{n,R}K_RX_{n,R}^\top\\
\mathbf{H}_n^R &:= \left((1-\delta_{i,j})\left(\widetilde{\mathbf{H}}_n^R\right)_{i,j}\right)_{1\leq i,j\leq n}.
\end{aligned}$$

We remark that $A_{n,R}^2 = I_R + E_{R,n}$ where $(E_{R,n})_{r,s} = (1/n)\sum_{i=1}^n \left(\phi_r(X_i)\phi_s(X_i) - \delta_{r,s}\right)$ for all $r, s \in [R]$. Denoting $\lambda(\mathbf{H}^R) = (\lambda_1,\dots,\lambda_R)$, we have

$$\begin{aligned}
\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2 \leq 4\big[&\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}^R))^2 + \delta_2(\lambda(\mathbf{H}^R),\lambda(\widetilde{\mathbf{H}}_n^R))^2\\
&+ \delta_2(\lambda(\widetilde{\mathbf{H}}_n^R),\lambda(\mathbf{H}_n^R))^2 + \delta_2(\lambda(\mathbf{H}_n^R),\lambda(\mathbf{H}_n))^2\big].
\end{aligned}$$

Bounding $\delta_2\left(\lambda(\mathbf{H}^R),\lambda(\widetilde{\mathbf{H}}_n^R)\right)^2$.
Let us consider some $\epsilon > 0$.
Using a singular value decomposition of $X_{n,R}$, one can show that $\lambda(X_{n,R}K_RX_{n,R}^\top) = \lambda(A_{n,R}K_RA_{n,R})$ which leads to

$$\begin{aligned}
\delta_2\left(\lambda(\mathbf{H}^R),\lambda(\widetilde{\mathbf{H}}_n^R)\right) &= \delta_2\left(\lambda(K_R),\lambda(X_{n,R}K_RX_{n,R}^\top)\right)\\
&= \delta_2\left(\lambda(K_R),\lambda(A_{n,R}K_RA_{n,R})\right)\\
&\leq \|K_R - A_{n,R}K_RA_{n,R}\|_F,
\end{aligned}$$

Using Equation (4.8) from Koltchinskii and Giné [2000, page 127], we get

$$\delta_2\left(\lambda(\mathbf{H}^R),\lambda(\widetilde{\mathbf{H}}_n^R)\right)^2 \leq 2\|K_R E_{R,n}\|_F^2 = 2\sum_{1\leq r,s\leq R}\lambda_s^2\left(\frac{1}{n}\sum_{i=1}^n \phi_r(X_i)\phi_s(X_i)-\delta_{r,s}\right)^2. \tag{5.16}$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(\delta_2\left(\lambda(\mathbf{H}^R),\lambda(\widetilde{\mathbf{H}}_n^R)\right)^2 \geq t\right)\\
\leq\ & \sum_{1\leq s,r\leq R}\mathbb{P}\left(\sqrt{2}|\lambda_s|\left|\frac{1}{n}\sum_{i=1}^n \phi_r(X_i)\phi_s(X_i)-\delta_{r,s}\right| \geq \sqrt{t}/R\right)\\
\leq\ & \sum_{1\leq s,r\leq R,\lambda_s\neq 0}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \phi_r(X_i)\phi_s(X_i)-\delta_{r,s}\right| \geq \sqrt{t}/(\sqrt{2}R|\lambda_s|)\right)\\
\leq\ & \sum_{1\leq s,r\leq R,\lambda_s\neq 0} 16\exp\left(-\left(Km^2\tau^2\right)^{-1}\frac{nt}{R^2|\lambda_s|^2\Upsilon_R^4}\right)\\
=\ & 16R^2\exp\left(-\left(Km^2\tau^2\right)^{-1}\frac{nt}{R^2\Lambda_R^2\Upsilon_R^4}\right),
\end{aligned}$$

where the last inequality follows from Proposition A.14 and where $K>0$ is a universal constant.

Bounding $\delta_2(\lambda(\widetilde{\mathbf{H}}_n^R),\lambda(\mathbf{H}_n^R))^2$.

$$\delta_2(\lambda(\widetilde{\mathbf{H}}_n^R),\lambda(\mathbf{H}_n^R))^2 \leq \|\widetilde{\mathbf{H}}_n^R-\mathbf{H}_n^R\|_F^2 = \frac{1}{n^2}\left(\sum_{i=1}^n h_R^2(X_i,X_i)\right) \leq \frac{\|h_R\|_\infty^2}{n} \tag{5.17}$$

Bounding $\delta_2(\lambda(\mathbf{H}_n^R),\lambda(\mathbf{H}_n))^2$.

$$\delta_2(\lambda(\mathbf{H}_n^R),\lambda(\mathbf{H}_n))^2 \leq \|\widetilde{\mathbf{H}}_n^R-\widetilde{\mathbf{H}}_n\|_F^2 = \frac{1}{n^2}\left(\sum_{1\leq i,j\leq n,\ i\neq j}(h-h_R)(X_i,X_j)^2\right). \tag{5.18}$$

Let us consider,

$$\forall x,y\in E,\quad m_R(x,y) := (h-h_R)^2(x,y)-s_R(x)-s_R(y)-\mathbb{E}_{\pi\otimes\pi}[(h-h_R)^2(X,Y)],$$

where $s_R(x)=\mathbb{E}_\pi[(h-h_R)^2(x,X)]-\mathbb{E}_{\pi\otimes\pi}[(h-h_R)^2(X,Y)]$. One can check that for any $x\in E$, $\mathbb{E}_\pi[m_R(x,X)]=\mathbb{E}_\pi[m_R(X,x)]=0$. Hence, m_R is π-canonical.

$$\begin{aligned}
&\frac{1}{n(n-1)}\left(\sum_{1\leq i,j\leq n,\ i\neq j}(h-h_R)(X_i,X_j)^2\right)\\
=\ & \frac{1}{n(n-1)}\sum_{1\leq i,j\leq n,\ i\neq j} m_R(X_i,X_j)+\frac{2}{n}\sum_{i=1}^n s_R(X_i)+\mathbb{E}_{\pi\otimes\pi}[(h-h_R)^2(X,Y)].
\end{aligned} \tag{5.19}$$

Using Theorem 5.1, we get that there exist two constants $\beta,\kappa>0$ such that for any $u\geq 1$, it holds with probability at least $1-\beta e^{-u}\log(n)$,

$$\frac{1}{n(n-1)}\sum_{1\leq i,j\leq n,\ i\neq j} m_R(X_i,X_j) \leq \kappa\|h-h_R\|_\infty \log n\left\{\frac{u}{n}\vee\left(\frac{u}{n}\right)^2\right\}.$$

Let us now consider some $t > 0$ such that

$$\kappa \|h - h_R\|_\infty \log n \left\{ \frac{u}{n} \vee \left(\frac{u}{n}\right)^2 \right\} \leq t. \tag{5.20}$$

The condition (5.20) is equivalent to

$$u \leq n \left\{ \frac{t}{\kappa \|h - h_R\|_\infty \log n} \wedge \left(\frac{t}{\kappa \|h - h_R\|_\infty \log n} \right)^{1/2} \right\},$$

which is satisfied in particular if t and u are such that

$$u = \frac{n}{\log n} \left\{ \left[\frac{t}{c}\right] \wedge \left[\frac{t}{c}\right]^{1/2} \right\},$$

where $c = \kappa \|h - h_R\|_\infty$. One can finally notice that for this choice of u, the condition $u \geq 1$ holds in particular for n large enough in order to have $n/\log n \geq \kappa \|h - h_R\|_\infty t^{-1}$.
We deduce from this analysis that for any $t > 0$, we have for n large enough to satisfy $n/\log n \geq \kappa \|h - h_R\|_\infty t^{-1}$,

$$\mathbb{P}\left(\frac{1}{n(n-1)} \sum_{1 \leq i,j \leq n,\ i \neq j} m_R(X_i, X_j) \geq t \right) \leq \beta \log(n) \exp\left(-\frac{n}{\log n} \left\{ \left[\frac{t}{c}\right] \wedge \left[\frac{t}{c}\right]^{1/2} \right\} \right).$$

Using the Hoeffding inequality from Proposition A.14, we get that for some universal constant $K > 0$,

$$\mathbb{P}\left(\frac{2}{n} \left| \sum_{i=1}^{n} s_R(X_i) \right| \geq t \right) \leq 16 \exp\left(-n \frac{t^2}{K m^2 \tau^2 \|h - h_R\|_\infty^2} \right).$$

We deduce that for some universal constant $K > 0$ it holds

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{n^2} \left(\sum_{1 \leq i,j \leq n,\ i \neq j} (h - h_R)(X_i, X_j)^2 \right) - \mathbb{E}_{\pi \otimes \pi}\left[(h - h_R)^2 \right] \geq t \right) \\ \leq\ & 16 \exp\left(-n \frac{t^2}{K m^2 \tau^2 \|h - h_R\|_\infty^2} \right) + \beta \log(n) \exp\left(-\frac{n}{4 \log n} \left\{ \left[\frac{t}{c}\right] \wedge \left[\frac{t}{c}\right]^{1/2} \right\} \right). \end{aligned}$$

Since $\mathbb{E}_{\pi \otimes \pi}\left[(h - h_R)^2(X, Y) \right] = \sum_{i > R, i \in I} \lambda_i^2$, we deduce that

$$\begin{aligned} &\mathbb{P}\left(\delta_2(\lambda(\mathbf{H}_n^R), \lambda(\mathbf{H}_n))^2 - \sum_{i > R, i \in I} \lambda_i^2 \geq t \right) \\ \leq\ & 16 \exp\left(-n \frac{t^2}{K m^2 \tau^2 \|h - h_R\|_\infty^2} \right) + \beta \log(n) \exp\left(-\frac{n}{4 \log n} \left\{ \left[\frac{t}{c}\right] \wedge \left[\frac{t}{c}\right]^{1/2} \right\} \right). \end{aligned}$$

Hence we proved that for any $u > 0$ such that $n/\log n \geq \kappa \|h - h_R\|_\infty u^{-1}$,

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{4} \delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 \geq \frac{\|h_R\|_\infty^2}{n} + 2 \sum_{i > R, i \in I} \lambda_i^2 + u \right) \\ \leq\ & 16 \exp\left(-n \frac{u^2}{K m^2 \tau^2 \|h - h_R\|_\infty^2} \right) + \beta \log(n) \exp\left(-\frac{n}{16 \log n} \left\{ \left[\frac{u}{c}\right] \wedge \left[\frac{u}{c}\right]^{1/2} \right\} \right) \\ +\ & 16 R^2 \exp\left(-\frac{nu}{K m^2 \tau^2 R^2 \Lambda_R^2 \Upsilon_R^2} \right). \end{aligned}$$

Considering $t>0$ and applying the previous inequality with $u=t+\frac{\kappa\|h-h_R\|_\infty \log n}{n}$, we get

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2 \geq \left(\|h_R\|_\infty^2+\kappa\|h-h_R\|_\infty^2\right)\frac{\log n}{n}+2\sum_{i>R,i\in I}\lambda_i^2+t\right)\\
\leq\ & 16\exp\left(-n\frac{t^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right)+\beta\log(n)\exp\left(-\frac{n}{16\log n}\left\{\left[\frac{t}{c}\right]\wedge\left[\frac{t}{c}\right]^{1/2}\right\}\right)\\
+\ & 16R^2\exp\left(-\frac{nt}{Km^2\tau^2R^2\Lambda_R^2\Upsilon_R^2}\right).
\end{aligned}$$

This concludes the proof of Theorem 5.15. □

5.6.2 Proof of Theorem 5.3.

We consider any $R\in\mathbb{N}^*$. We remark that for any $x,y\in E$,

$$\begin{aligned}
|h_R(x,y)| &= \left|\sum_{r=1}^R \lambda_r\phi_r(x)\phi_r(y)\right|\\
&\leq \left(\sum_{r=1}^R|\lambda_r|\phi_r(x)^2\right)^{1/2}\times\left(\sum_{r=1}^R|\lambda_r|\phi_r(y)^2\right)^{1/2} \text{ (using Cauchy-Schwarz inequality)}\\
&\leq S,
\end{aligned}$$

which proves that $\|h_R\|_\infty\leq S$. Similar computations lead to $\|h-h_R\|_\infty\leq S$.
Using Theorem 5.15 we get for any $t>0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2 \geq \frac{S^2(1+\kappa)\log n}{n}+2\sum_{i>R,i\in I}\lambda_i^2+t\right)\\
\leq\ & 16\exp\left(-n\frac{t^2}{Km^2\tau^2S^2}\right)+\beta\log(n)\exp\left(-\frac{n}{16\log n}\left\{\left[\frac{t}{\kappa S}\right]\wedge\left[\frac{t}{\kappa S}\right]^{1/2}\right\}\right)\\
+\ & 16R^2\exp\left(-\frac{nt}{Km^2\tau^2R^2\Lambda^2\Upsilon^2}\right),
\end{aligned}$$

where $\Lambda:=\sup_{r\geq 1}|\lambda_r|<\infty$. Choosing $R^2=\lceil\sqrt{n}\rceil$, we get

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2 \geq \frac{S^2(1+\kappa)\log n}{n}+2\sum_{i>\lceil n^{1/4}\rceil,i\in I}\lambda_i^2+t\right)\\
\leq\ & 32\sqrt{n}\exp\left(-\mathcal{C}\min\left(nt^2,\sqrt{n}t\right)\right)+\beta\log(n)\exp\left(-\frac{n}{\log n}\min\left(\mathcal{B}t,(\mathcal{B}t)^{1/2}\right)\right),
\end{aligned}$$

where $\mathcal{B}=(K\kappa S)^{-1}$ and $\mathcal{C}=K^{-1}\left(m^2\tau^2(S+\Lambda\Upsilon)\right)^{-2}$.

5.7 Proofs for Section 5.4

In this section, for any $k\geq 0$ we denote $\mathbb{E}_k$ the conditional expectation with respect to the σ-algebra $\sigma(X_1,\dots,X_k)$.

5.7.1 Proof of Theorem 5.7

By definition of $\mathcal{M}^n$, we want to bound

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}M_t\geq\epsilon\right),$$

which takes the form

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathcal{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[M_t]]-\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[M_t-\mathbb{E}_{t-b_n}[M_t]]\geq\epsilon\right)\\ \leq\;&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathcal{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[M_t]]\geq\epsilon/2\right)+\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathbb{E}_{t-b_n}[M_t]-M_t]\geq\epsilon/2\right).\end{aligned}\tag{5.21}$$

5.7.1.1 Step 1: Martingale difference

We first deal with the second term of Eq.(5.21). Note that we can write

$$\sum_{t=c_n}^{n-1}[\mathbb{E}_{t-b_n}[M_t]-M_t]=\sum_{t=c_n}^{n-1}\sum_{k=1}^{b_n}[\mathbb{E}_{t-k}[M_t]-\mathbb{E}_{t-k+1}[M_t]]=\sum_{k=1}^{b_n}\sum_{t=c_n}^{n-1}[\mathbb{E}_{t-k}[M_t]-\mathbb{E}_{t-k+1}[M_t]].$$

Let us consider some $k\in\{1,\dots,b_n\}$, then we have that $V_t^{(k)}=(\mathbb{E}_{t-k}[M_t]-\mathbb{E}_{t-k+1}[M_t])/(n-c_n)$ is a martingale difference sequence, i.e. $\mathbb{E}_{t-k}[V_t^{(k)}]=0$. Since the loss function is bounded in $[0,1]$, we have $|V_t^{(k)}|\leq 2/(n-c_n),\ t=1,\dots,n$. Therefore by the Hoeffding-Azuma inequality, $\sum_t V_t^{(k)}$ can be bounded such that

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathbb{E}_{t-k}[M_t]-\mathbb{E}_{t-k+1}[M_t]]\geq\frac{\epsilon}{2b_n}\right)\leq\exp\left(-\frac{(1-c)n\epsilon^2}{8b_n^2}\right).$$

We deduce that

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathbb{E}_{t-b_n}[M_t]-M_t]\geq\epsilon/2\right)\leq b_n\exp\left(-\frac{(1-c)n\epsilon^2}{8b_n^2}\right).\tag{5.22}$$

5.7.1.2 Step 2: Symmetrization by a ghost sample

In this step we bound the first term in Eq.(5.21). Let us start by introducing a ghost sample $\{\xi_j\}_{1\leq j\leq n}$, where the random variables ξ_j are i.i.d with distribution π. Recall the definition of M_t and define $\widetilde{M}_t$ as

$$M_t=\frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\ell(h_{t-b_n},X_t,X_i),\qquad\widetilde{M}_t=\frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\ell(h_{t-b_n},X_t,\xi_i).$$

The difference between $\widetilde{M}_t$ and M_t is that M_t is the sum of the loss incurred by h_{t-b_n} on the current instance X_t and all the previous examples $X_j,\ j=1,\dots,t-b_n$ on which h_{t-b_n} is trained, while $\widetilde{M}_t$ is the loss incurred by the same hypothesis h_{t-b_n} on the current instance X_t and an independent set of examples $\xi_j,\ j=1,\dots,t-b_n$.
First remark that we have

$$\begin{aligned}&\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathcal{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[M_t]]\\ =\;&\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathcal{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[\widetilde{M}_t]\right]+\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_{t-b_n}[\widetilde{M}_t]-\mathbb{E}_{t-b_n}[M_t]\right].\end{aligned}\tag{5.23}$$

The first term of Eq.(5.23) is handled in Wang et al. [2012] by relying heavily on the assumption that samples are i.i.d [see Wang et al., 2012, Claim 1]. Hence, the approach of Wang and al. cannot be adapted in our framework. To overcome this difficulty, we use the uniform ergodicity of the Markov chain. This is where the use of the burning parameter b_n is essential.

Since ℓ is in $[0,1]$, the first term can be bounded directly using the uniform ergodicity of the Markov chain $(X_i)_i$ as follows

$$\begin{aligned}
& \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathcal{R}(h_{t-b_n}) - \mathbb{E}_{t-b_n}[\widetilde{M}_t]\right] \\
= & \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\int_{x\in E}\left(d\pi(x)\mathbb{E}_{X\sim\pi}[\ell(h_{t-b_n},x,X)] - P^{b_n}(X_{t-b_n},dx)\mathbb{E}_{X\sim\pi}[\ell(h_{t-b_n},x,X)]\right) \\
= & \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\int_{x\in E}\mathbb{E}_{X\sim\pi}[\ell(h_{t-b_n},x,X)]\left(d\pi(x) - P^{b_n}(X_{t-b_n},dx)\right) \\
\leq & \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\int_{x\in E}\left|d\pi(x) - P^{b_n}(X_{t-b_n},dx)\right| \\
\leq & L\rho^{b_n},
\end{aligned}$$

where we used Eq.(5.6). It remains to control

$$\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t]\right],$$

and we follow an approach similar to Wang et al. [2012]. Let us remind that M_t and $\widetilde{M}_t$ depend on the hypothesis h_{t-b_n} and let us define $L_t(h_{t-b_n}) = \left[\mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t]\right]$. We have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}L_t(h_{t-b_n})\geq\epsilon\right) &\leq \mathbb{P}\left(\sup_{\widehat{h}_{c_n-b_n},\ldots,\widehat{h}_{n-1-b_n}}\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}L_t(\widehat{h}_{t-b_n})\geq\epsilon\right) \\
&\leq \sum_{t=c_n}^{n-1}\mathbb{P}\left(\sup_{\widehat{h}\in\mathcal{H}}L_t(\widehat{h})\geq\epsilon\right). \qquad (5.24)
\end{aligned}$$

To bound the right hand side of Eq.(5.24) we give first the following Lemma.

Lemma 5.16. *Given any function $f\in\mathcal{H}$ and any $t\geq c_n$,*

$$\forall\epsilon>0,\quad \mathbb{P}\left(L_t(f)\geq\epsilon\right)\leq 16\exp\left(-(t-b_n)C(m,\tau)\epsilon^2\right).$$

In the i.i.d. framework, the counterpart of Lemma 5.16 follows from a straightforward application of McDiarmid's inequality [see Wang et al., 2012, Lemma 5]. In our work, we consider uniformly ergodic Markov chains and the proof of Lemma 5.16 requires extra work. We apply a concentration inequality for Markov chains (see Proposition A.14) which needs to hold for any initial distribution. We apply Proposition A.14 by considering the time-reversed sequence and this is where we use the reversibility of the chain.

Proof of Lemma 5.16. Note that

$$\begin{aligned}
L_t(f) &= \mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t] \\
&= \frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\left(\mathbb{E}_{t-b_n}[\ell(f,X_t,\xi_i)] - \mathbb{E}_{t-b_n}[\ell(f,X_t,X_i)]\right) \\
&= \frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\mathbb{E}_{\xi\sim\pi}\left[\mathbb{E}_{X_t\sim P^{b_n}(X_{t-b_n},\cdot)}\{\ell(f,X_t,\xi)\}\right] - \mathbb{E}_{X_t\sim P^{b_n}(X_{t-b_n},\cdot)}\{\ell(f,X_t,X_i)\}.
\end{aligned}$$

Hence, denoting $m(f, X_{t-b_n}, x) = \mathbb{E}_{X_t \sim P^{b_n}(X_{t-b_n}, \cdot)}\{\ell(f, X_t, x)\}$, we get

$$L_t(f) \leq \frac{1}{t-b_n} \sum_{i=1}^{t-b_n} \left\{ \mathbb{E}_{\xi \sim \pi} \left[m(f, X_{t-b_n}, \xi) \right] - m(f, X_{t-b_n}, X_i) \right\}.$$

By the reversibility of the chain $(X_i)_{i \geq 1}$, we know that the sequence $(X_{t-b_n}, X_{t-b_n-1}, \ldots, X_1)$ conditionally on X_{t-b_n} is a Markov chain with stationary distribution π. Applying Proposition A.14, we get that

$$\begin{aligned} & \mathbb{P}\left(L_t(f) \geq \epsilon \mid X_{t-b_n}\right) \\ \leq \; & \mathbb{P}\left(\frac{1}{t-b_n} \sum_{i=1}^{t-b_n} \left\{ \mathbb{E}_{\xi_i \sim \pi} \left[m(f, X_{t-b_n}, \xi_i) \right] - m(f, X_{t-b_n}, X_i) \right\} \geq \epsilon \mid X_{t-b_n} \right) \\ \leq \; & 16 \exp\left(-(t-b_n) C(m,\tau) \epsilon^2 \right), \end{aligned}$$

for some constant $C(m, \tau) > 0$ depending only on m and τ. Then we deduce that

$$\begin{aligned} \mathbb{P}\left(L_t(f) \geq \epsilon\right) &= \mathbb{E}\left[\mathbb{E}\left\{ \mathbb{1}_{L_t(f) \geq \epsilon} \mid X_{t-b_n} \right\} \right] \\ &= \mathbb{E}\left[\mathbb{P}\left\{ L_t(f) \geq \epsilon \mid X_{t-b_n} \right\} \right] \\ &\leq 16 \exp\left(-(t-b_n) C(m,\tau) \epsilon^2 \right), \end{aligned}$$

which concludes the proof of Lemma 5.16. □

The following two Lemmas are key elements to prove Lemma 5.19. Their proofs are strictly analogous to the proofs of Lemmas 6, 7 and 8 from Wang et al. [2012].

Lemma 5.17. *[cf. Wang et al., 2012, Lemma 6] For any two functions $h_1, h_2 \in \mathcal{H}$, the following equation holds*

$$|L_t(h_1) - L_t(h_2)| \leq 2\mathrm{Lip}(\phi) \|h_1 - h_2\|_\infty.$$

Lemma 5.18. *Let $\mathcal{H} = S_1 \cup \cdots \cup S_l$ and $\epsilon > 0$. Then*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} L_t(h) \geq \epsilon \right) \leq \sum_{j=1}^{l} \mathbb{P}\left(\sup_{h \in S_j} L_t(h) \geq \epsilon \right).$$

Lemma 5.19. *[cf. Wang et al., 2012, Lemma 6] For any $c_n \leq t \leq n$, it holds*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} L_t(h) \geq \epsilon \right) \leq 16 \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{4\mathrm{Lip}(\phi)} \right) \exp\left(-\frac{(t-b_n) C(m,\tau) \epsilon^2}{4} \right).$$

Combining Lemma 5.19 and Eq.(5.24), we have

$$\mathbb{P}\left(\frac{1}{n-c_n} \sum_{t=c_n}^{n-1} L_t(h_{t-b_n}) \geq \epsilon \right) \leq 16 \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{4\mathrm{Lip}(\phi)} \right) n \exp\left(-\frac{(c_n-b_n) C(m,\tau) \epsilon^2}{4} \right).$$

We deduce that

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{n-c_n} \sum_{t=c_n}^{n-1} \left[\mathcal{R}(h_{t-b_n}) - \mathbb{E}_{t-b_n}[M_t] \right] \geq \epsilon/2 \right) \\ \leq \; & \mathbb{P}\left(L\rho^{b_n} + \frac{1}{n-c_n} \sum_{t=c_n}^{n-1} \left[\mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t] \right] \geq \epsilon/2 \right) \\ \leq \; & 16 \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8\mathrm{Lip}(\phi)} \right) n \exp\left(-\frac{(c_n-b_n) C(m,\tau) \left(\epsilon/2 - L\rho^{b_n}\right)^2}{4} \right). \end{aligned}$$

5.7.1.3 Step 3: Conclusion of the proof

By considering dependent random variables, we needed the introduction of the burning parameter b_n (see Eq.(5.23)). This situation brings extra technicalities to conclude the proof.

From the previous inequality and (5.22), we get

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}M_t\geq\epsilon\right)\\ \leq\;& b_n\exp\left(-\frac{(1-c)n\epsilon^2}{8b_n^2}\right)+16\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8\mathrm{Lip}(\phi)}\right)n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\left(\epsilon/2-L\rho^{b_n}\right)^2}{4}\right).\end{aligned}$$

Note that $(c_n-b_n)\epsilon\rho^{b_n}\underset{n\to\infty}{=}o\left(n\epsilon n^{q\log(\rho)}\right)\underset{n\to\infty}{=}o\left(n^{1+\xi+q\log(\rho)}\right)$ because by assumption $\epsilon\underset{n\to\infty}{=}o\left(n^{\xi}\right)$. However, by choice of q we have

$$1+\xi+q\log(\rho)=1+\xi+\frac{1+\xi}{\log(1/\rho)}\log(\rho)=0,$$

and we finally get that $(c_n-b_n)\epsilon\rho^{b_n}\underset{n\to\infty}{=}o\left(1\right)$. We deduce that for n large enough it holds

$$\exp\left(-\frac{(c_n-b_n)C(m,\tau)\left(\epsilon/2-L\rho^{b_n}\right)^2}{4}\right)\leq 2\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16}\right).$$

Then, noticing that

$$\exp\left(-\frac{(1-c)n\epsilon^2}{8b_n^2}\right)\underset{n\to\infty}{=}\mathcal{O}\left(\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16b_n^2}\right)\right),$$

we finally get for n large enough

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}M_t\geq\epsilon\right)\\ \leq\;&\left[32\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8\mathrm{Lip}(\phi)}\right)+1\right]b_n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16b_n^2}\right).\end{aligned}$$

5.7.2 Proof of Theorem 5.9

Theorem 5.7 shows that

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\mathcal{M}^n\right|\geq\epsilon\right)\\ \leq\;&\left[32\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8\mathrm{Lip}(\phi)}\right)+1\right]b_n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16b_n^2}\right),\end{aligned}\tag{5.25}$$

and the assumption on the space $\mathcal{H}$ gives that for some $\theta>0$, it holds for any $\eta>0$, $\log\mathcal{N}(\mathcal{H},\eta)=\mathcal{O}(\eta^{-\theta})$. By taking $\epsilon=\frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}}$ it is straightforward to prove that the logarithm of the right hand side of Eq.(5.25) goes to $-\infty$ as $n\to+\infty$. This concludes the proof of the first part of Theorem 5.9. Since the result from Theorem 5.7 trivially holds when considering $h_1=\cdots=h_{n-1}=h^*$, the previous computations show that for any $\delta>0$ there exists some $N\in\mathbb{N}$ such that for any $n\geq N$ it holds with

probability at least $1-\delta$,

$$\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\mathcal{M}^n\right| \vee |\mathcal{M}^n(h^*,\dots,h^*)-\mathcal{M}^n| \leq \frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}}.$$

Hence, by considering that the online learner has a regret bound $\mathfrak{R}_n$ (cf. Definition 5.8), we get that for any $\delta>0$ there exists some $N\in\mathbb{N}$ such that for any $n\geq N$ it holds with probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\mathcal{R}(h^*)\\
\leq\ &\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-b_n})-\mathcal{M}^n+\mathcal{M}^n-\mathcal{M}^n(h^*,\dots,h^*)+\mathcal{M}^n(h^*,\dots,h^*)-\mathcal{R}(h^*)\\
\leq\ &2\frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}}+\mathcal{M}^n-\inf_{h\in\mathcal{H}}\mathcal{M}^n(h,\dots,h)\leq 2\frac{\log(n)\log(\log n)}{n^{\frac{1}{2+\theta}}}+\mathfrak{R}_n,
\end{aligned}$$

which concludes the proof of Theorem 5.9.

5.7.3 Proof of Theorem 5.10

The proof of Theorem 5.10 has two main steps. First, we show that $\mathcal{R}(\widehat{h})$ is close to $\min_{c_n\leq t\leq n-1}\mathcal{R}(h_{t-b_n})+2c_\gamma(n-t)$ with high probability. Then we show that $\min_{c_n\leq t\leq n-1}\mathcal{R}(h_{t-b_n})+2c_\gamma(n-t)$ is close to $\mathcal{M}^n$ with high probability. The second step is similar to the proof of Wang et al. [2012]. For the first step, we need a concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. This is where we use the Hoeffding decomposition and Theorem 5.1 (see Section 5.2.2).

Let us recall that for any $1\leq t\leq n-2$, $\widehat{\mathcal{R}}(h_{t-b_n},t+1)=\binom{n-t}{2}^{-1}\sum_{k>i,i\geq t+1}^{n}\ell(h_{t-b_n},X_i,X_k)$. We define

$$\ell(h,x):=\mathbb{E}_\pi[\ell(h,X,x)]-\mathcal{R}(h),\ \text{and}\ \widetilde{\ell}(h,x,y)=\ell(h,x,y)-\ell(h,x)-\ell(h,y)-\mathcal{R}(h).$$

Then for any $t\in\{b_n+1,\dots,n-2\}$ we have the following decomposition

$$\widehat{\mathcal{R}}(h_{t-b_n},t+1)-\mathcal{R}(h_{t-b_n})=\binom{n-t}{2}^{-1}\sum_{k>i,i\geq t+1}^{n}\widetilde{\ell}(h_{t-b_n},X_i,X_k)+\frac{2}{n-t}\sum_{i=t+1}^{n}\ell(h_{t-b_n},X_i). \quad (5.26)$$

One can check that for any $x\in E$, $\mathbb{E}_\pi\left[\widetilde{\ell}(h,X,x)\right]=\mathbb{E}_\pi\left[\widetilde{\ell}(h,x,X)\right]=0$. Moreover, for any hypothesis $h\in\mathcal{H}$, $\|\widetilde{\ell}(h,\cdot,\cdot)\|_\infty\leq 4$ (because the loss function ℓ takes its value in $[0,1]$). Hence, for any fixed hypothesis $h\in\mathcal{H}$, the kernel $\widetilde{\ell}(h,\cdot,\cdot)$ satisfies Assumption 3. Applying Theorem 5.1, we know that there exist constants $\beta,\kappa>0$ such that for any $t\in\{b_n+1,\dots,n-2\}$ and for any $\gamma\in(0,1)$, it holds with probability at least $1-\gamma$,

$$\left|\binom{n-t}{2}^{-1}\sum_{k>i,i\geq t+1}^{n}\widetilde{\ell}(h_{t-b_n},X_i,X_k)\right|\leq\kappa\frac{\log(n-t-1)}{n-t-1}\log((\beta\vee e^1)\log(n-t+1)/\gamma)^2.$$

Note that we used that for $u=\log\left((\beta\vee e^1)\log(n-t+1)/\gamma\right)\geq 1$ it holds

$$\log n\left\{\frac{u}{n}\vee\left(\frac{u}{n}\right)^2\right\}\leq\frac{\log n}{n}u^2.$$

Using Proposition A.14, we also have that for any $t\in\{b_n+1,\dots,n-2\}$ and any $\epsilon>0$,

$$\mathbb{P}\left(\left|\frac{2}{n-t}\sum_{i=t+1}^{n}\ell(h_{t-b_n},X_i)\right|>\epsilon\right)\leq 32\exp\left(-C(m,\tau)(n-t)\epsilon^2\right),$$

where $C(m,\tau) = (Km^2\tau^2)^{-1} > 0$ for some universal constant K (one can check from the proof of Proposition A.14 that $K = 7 \times 10^3$ fits). We get that for any $t \in \{b_n + 1, \dots, n-2\}$ and any $\gamma \in (0,1)$, it holds with probability at least $1-\gamma$,

$$\left| \frac{2}{n-t} \sum_{i=t+1}^{n} \ell(h_{t-b_n}, X_i) \right| \leq \frac{\log(32/\gamma)^{1/2} C(m,\tau)^{-1/2}}{\sqrt{n-t}}.$$

We deduce that for any $t \in \{b_n + 1, \dots, n-2\}$ and any fixed $\gamma \in (0,1)$, it holds with probability at least $1-\gamma$,

$$\left| \widehat{\mathcal{R}}(h_{t-b_n}, t+1) - \mathcal{R}(h_{t-b_n}) \right| \leq C(m,\tau)^{-1/2} \sqrt{\frac{\log(64/\gamma)}{n-t}},$$

i.e.

$$\mathbb{P}\left(\left| \widehat{\mathcal{R}}(h_{t-b_n}, t+1) - \mathcal{R}(h_{t-b_n}) \right| \geq c_\gamma (n-t) \right) \leq \frac{\gamma}{(n-c_n)(n-c_n+1)}. \tag{5.27}$$

Based on the selection procedure of the hypothesis $\widehat{h}$ defined in Eq.(5.13), the concentration result Eq.(5.27) allows us to show that $\mathcal{R}(\widehat{h})$ is close to $\min_{c_n \leq t \leq n-1} \mathcal{R}(h_{t-b_n}) + 2c_\gamma(n-t)$ with high probability. This is stated by Lemma 5.20 which is proved in Section 5.7.4.

Lemma 5.20. *Let $h_0, \dots, h_{n-1}$ be the set of hypotheses generated by an arbitrary online algorithm $\mathcal{A}$ working with a pairwise loss ℓ which satisfies the conditions given in Theorem 5.7. Then for any $\gamma \in (0,1)$, we have*

$$\mathbb{P}\left(\mathcal{R}(\widehat{h}) > \min_{c_n \leq t < n-1} (\mathcal{R}(h_{t-b_n}) + 2c_\gamma(n-t)) \right) \leq \gamma.$$

To conclude the proof, we need to show that $\min_{c_n \leq t \leq n-1} \mathcal{R}(h_{t-b_n}) + 2c_\gamma(n-t)$ is close to $\mathcal{M}^n$.

First we remark that

$$\begin{aligned}
& \min_{c_n \leq t \leq n-1} \mathcal{R}(h_{t-b_n}) + 2c_\gamma(n-t) \\
= & \min_{c_n \leq t \leq n-1} \min_{t \leq i \leq n-1} \mathcal{R}(h_{i-b_n}) + 2c_\gamma(n-i) \\
\leq & \min_{c_n \leq t \leq n-1} \frac{1}{n-t} \sum_{i=t}^{n-1} (\mathcal{R}(h_{i-b_n}) + 2c_\gamma(n-i)) \\
\leq & \min_{c_n \leq t \leq n-1} \left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathcal{R}(h_{i-b_n}) + \frac{2}{n-t} \sum_{i=t}^{n-1} \sqrt{\frac{C(m,\tau)^{-1}}{n-i} \log \frac{64(n-c_n)(n-c_n+1)}{\gamma}} \right) \\
\leq & \min_{c_n \leq t \leq n-1} \left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathcal{R}(h_{i-b_n}) + \frac{2}{n-t} \sum_{i=t}^{n-1} \sqrt{\frac{2C(m,\tau)^{-1}}{n-i} \log \frac{64(n-c_n+1)}{\gamma}} \right) \\
\leq & \min_{c_n \leq t \leq n-1} \left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathcal{R}(h_{i-b_n}) + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t} \log \frac{64(n-c_n+1)}{\gamma}} \right),
\end{aligned}$$

where the last inequality holds because $\sum_{i=1}^{n-t} \sqrt{1/i} \leq 2\sqrt{n-t}$. Indeed, $x \mapsto 1/\sqrt{x}$ is a decreasing and continuous function and a classical serie/integral approach leads to

$$\sum_{i=1}^{n-t} \sqrt{1/i} \leq 1 + \int_1^{n-t} \frac{1}{\sqrt{x}} dx = 1 + \left[2\sqrt{x}\right]_1^{n-t} \leq 2\sqrt{n-t}.$$

We define $\mathcal{M}_t^n := \frac{1}{n-t} \sum_{m=t}^{n-1} M_m$. From Theorem 5.7, one can see that for each $t = c_n, \dots, n-1$,

$$\mathbb{P}\left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathcal{R}(h_{i-b_n}) \geq \mathcal{M}_t^n + \epsilon \right) \leq \left[32\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8\mathrm{Lip}(\phi)} \right) + 1 \right] b_n \exp\left(-\frac{(t-b_n)C(m,\tau)\epsilon^2}{16 b_n^2} \right).$$

Let us set

$$K_t = \mathcal{M}_t^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t}\log\frac{64(n-c_n+1)}{\gamma}} + \epsilon.$$

Using the fact that if $\min(a_1,a_2) \leq \min(b_1,b_2)$ then either $a_1 \leq b_1$ or $a_2 \leq b_2$, we can write

$$\begin{aligned}
&\mathbb{P}\left(\min_{c_n\leq t\leq n-1}\mathcal{R}(h_{t-b_n}) + 2c_\gamma(n-t) \geq \min_{c_n\leq t\leq n-1} K_t\right)\\
\leq\ &\mathbb{P}\left(\min_{c_n\leq t\leq n-1}\left(\frac{1}{n-t}\sum_{i=t}^{n-1}\mathcal{R}(h_{i-b_n}) + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t}\log\frac{64(n-c_n+1)}{\gamma}}\right) \geq \min_{c_n\leq t\leq n-1} K_t\right)\\
\leq\ &\sum_{t=c_n}^{n-1}\mathbb{P}\left(\frac{1}{n-t}\sum_{i=t}^{n-1}\mathcal{R}(h_{i-b_n}) + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t}\log\frac{64(n-c_n+1)}{\gamma}} \geq K_t\right)\\
=\ &\sum_{t=c_n}^{n-1}\mathbb{P}\left(\frac{1}{n-t}\sum_{i=t}^{n-1}\mathcal{R}(h_{i-b_n}) \geq \mathcal{M}_t^n + \epsilon\right)\\
\leq\ &(n-c_n)\left[32\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8Lip(\phi)}\right)+1\right]b_n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16b_n^2}\right)\\
\leq\ &\left[32\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8Lip(\phi)}\right)+1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16b_n^2}+2\log n\right).
\end{aligned}$$

Using Lemma 5.20, we get

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h}) \geq \min_{c_n\leq t\leq n-1}\mathcal{M}_t^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t}\log\frac{64(n-c_n+1)}{\gamma}} + \epsilon\right)\\
\leq\ &\gamma + \left[32\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8Lip(\phi)}\right)+1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16b_n^2}+2\log n\right),
\end{aligned}$$

which gives in particular

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h}) \geq \mathcal{M}^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-c_n}\log\frac{64(n-c_n+1)}{\gamma}} + \epsilon\right)\\
\leq\ &\gamma + \left[32\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8Lip(\phi)}\right)+1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{16b_n^2}+2\log n\right).
\end{aligned}$$

We substitute ϵ with $\epsilon/2$ and we choose γ such that $4\sqrt{\frac{2C(m,\tau)^{-1}}{n-c_n}\log\frac{64(n-c_n+1)}{\gamma}} = \epsilon/2$ with n large enough to ensure that $\gamma < 1$. We have for any $c > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h}) \geq \mathcal{M}^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-c_n}\log\frac{64(n-c_n+1)}{\gamma}} + \frac{\epsilon}{2}\right)\\
\leq\ &64(n-c_n+1)\exp\left(-\frac{(n-c_n)C(m,\tau)\epsilon^2}{128}\right) + \left[32\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{16Lip(\phi)}\right)+1\right]\times\\
&\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{(16b_n)^2}+2\log n\right)\\
\leq\ &32\left[\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{16Lip(\phi)}\right)+1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\epsilon^2}{(16b_n)^2}+2\log n\right),
\end{aligned}$$

where these inequalities hold for n large enough.

5.7.4 Proof of Lemma 5.20

Let

$$T^* := \arg \min_{c_n \le t < n-1} (\mathcal{R}(h_{t-b_n}) + 2c_\gamma(n-t)),$$

and $h^* = h_{T^*-b_n}$ is the corresponding hypothesis that minimizes the penalized true risk and let $\widehat{\mathcal{R}}^* = \widehat{\mathcal{R}}(h^*, T^*+1)$ to be the penalized empirical risk of $h_{T^*-b_n}$. Set, for brevity

$$\widehat{\mathcal{R}}_{t-b_n} = \widehat{\mathcal{R}}(h_{t-b_n}, t+1),$$

and let

$$\widehat{T} := \arg \min_{c_n \le t < n-1} (\widehat{\mathcal{R}}_{t-b_n} + c_\gamma(n-t)),$$

where $\widehat{h}$ coincides with $h_{\widehat{T}-b_n}$. Using this notation and since

$$\widehat{\mathcal{R}}_{\widehat{T}-b_n} + c_\gamma(n-\widehat{T}) \le \widehat{\mathcal{R}}^* + c_\gamma(n-T^*),$$

holds with certainty, we have

$$\begin{aligned}
& \mathbb{P}\left(\mathcal{R}(\widehat{h}) > \mathcal{R}(h^*) + \mathcal{E}\right) \\
= \; & \mathbb{P}\left(\mathcal{R}(\widehat{h}) > \mathcal{R}(h^*) + \mathcal{E}, \widehat{\mathcal{R}}_{\widehat{T}-b_n} + c_\gamma(n-\widehat{T}) \le \widehat{\mathcal{R}}^* + c_\gamma(n-T^*)\right) \\
\le \; & \mathbb{P}\left(\bigcup_{c_n \le t \le n-1} \left\{\mathcal{R}(h_{t-b_n}) > \mathcal{R}(h^*) + \mathcal{E}, \widehat{\mathcal{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathcal{R}}^* + c_\gamma(n-T^*)\right\}\right) \\
\le \; & \sum_{t=c_n}^{n-1} \mathbb{P}\left(\mathcal{R}(h_{t-b_n}) > \mathcal{R}(h^*) + \mathcal{E}, \widehat{\mathcal{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathcal{R}}^* + c_\gamma(n-T^*)\right),
\end{aligned}$$

where $\mathcal{E}$ is a positive-valued random variable to be specified. Now we remark that if

$$\widehat{\mathcal{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathcal{R}}^* + c_\gamma(n-T^*), \tag{5.28}$$

holds, then at least one of the following three conditions must hold

$$\begin{array}{lcl}
(i) & \widehat{\mathcal{R}}_{t-b_n} & \le \mathcal{R}(h_{t-b_n}) - c_\gamma(n-t) \\
(ii) & \widehat{\mathcal{R}}^* & > \mathcal{R}(h^*) + c_\gamma(n-T^*) \\
(iii) & \mathcal{R}(h_{t-b_n}) - \mathcal{R}(h^*) & \le 2c_\gamma(n-T^*).
\end{array}$$

Stated otherwise, if Eq.(5.28) holds for some $t \in \{c_n, \dots, n-1\}$ then

- either $t = T^*$ and (iii) holds trivially.
- or $t \neq T^*$ which can occur because
  - $\widehat{\mathcal{R}}_{t-b_n}$ underestimates $\mathcal{R}(h_{t-b_n})$ and (i) holds.
  - $\widehat{\mathcal{R}}^*$ overestimates $\mathcal{R}(h^*)$ and (ii) holds.
  - n is too small to statistically distinguish $\mathcal{R}(h_{t-b_n})$ and $\mathcal{R}(h^*)$, and (iii) holds.

Therefore, for any fixed t, we have

$$\begin{aligned}
& \mathbb{P}\left(\mathcal{R}(h_{t-b_n}) > \mathcal{R}(h^*) + \mathcal{E}, \widehat{\mathcal{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathcal{R}}^* + c_\gamma(n-T^*)\right) \\
\le \; & \mathbb{P}\left(\widehat{\mathcal{R}}_{t-b_n} \le \mathcal{R}(h_{t-b_n}) - c_\gamma(n-t)\right) + \mathbb{P}\left(\widehat{\mathcal{R}}^* > \mathcal{R}(h^*) + c_\gamma(n-T^*)\right) \\
& + \mathbb{P}\left(\mathcal{R}(h_{t-b_n}) - \mathcal{R}(h^*) \le 2c_\gamma(n-T^*)\,,\; \mathcal{R}(h_{t-b_n}) > \mathcal{R}(h^*) + \mathcal{E}\right).
\end{aligned}$$

By choosing $\mathcal{E} = 2c_\gamma(n - T^*)$, the last term in the previous inequality is zero and we can write

$$\begin{aligned}
& \mathbb{P}\left(\mathcal{R}(\widehat{h}) > \mathcal{R}(h^*) + 2c_\gamma(n - T^*)\right) \\
\leq & \sum_{t=c_n}^{n-1} \mathbb{P}\left(\widehat{\mathcal{R}}_{t-b_n} \leq \mathcal{R}(h_{t-b_n}) - c_\gamma(n-t)\right) + (n - c_n)\mathbb{P}\left(\widehat{\mathcal{R}}^* > \mathcal{R}(h^*) + c_\gamma(n - T^*)\right) \\
\leq & \frac{(n-c_n)\gamma}{(n-c_n)(n-c_n+1)} + (n-c_n)\left\{\sum_{t=c_n}^{n-1} \mathbb{P}\left(\widehat{\mathcal{R}}_{t-b_n} > \mathcal{R}(h_{t-b_n}) + c_\gamma(n-t)\right)\right\} \text{ (using (5.27))} \\
\leq & \frac{\gamma}{n-c_n+1} + (n-c_n)^2 \frac{\gamma}{(n-c_n)(n-c_n+1)} \text{ (using Eq.(5.27))} \\
\leq & \frac{\gamma}{n-c_n+1} + (n-c_n)\frac{\gamma}{n-c_n+1} = \gamma.
\end{aligned}$$

5.7.5 Proof of Corollary 5.11

The proof of Corollary 5.11 is analogous to the proof of Theorem 5.9 by applying Theorem 5.10 (instead of Theorem 5.7) and by choosing $\epsilon = \frac{\log^2 n}{n^{\frac{1}{2+\theta}}}$.

5.8 Proofs for Section 5.5

5.8.1 Proof of Theorem 5.12

In the following, $\mathbb{P}_g$ will denote the distribution of the Markov chain if the stationary distribution of the chain is assumed to have a density g with respect to the Lebesgue measure on $\mathbb{R}$. We consider $q = q_1 \vee q_2$ where $q_1, q_2 \in [1, \infty)$ are such that $\frac{1}{p_1} + \frac{1}{q_1} = 1$ and $\frac{1}{p_2} + \frac{1}{q_2} = 1$.

The main tool of the proof is the Hoeffding (also called canonical) decomposition of the U-statistic $\widehat{\theta}_m$. We introduce the processes U_n and P_n defined by

$$U_n(h) = \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} h(X_i, X_j), \quad P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

We also define $P(h) = \langle h, f \rangle$. By setting, for all $m \in \mathcal{M}$,

$$H_m(x, y) = \sum_{l \in \mathcal{L}_m} (p_l(x) - a_l)(p_l(y) - a_l),$$

with $a_l = \langle f, p_l \rangle$, we obtain the decomposition

$$\widehat{\theta}_m = U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f)) + \|\Pi_{S_m}(f)\|_2^2.$$

Let us consider β in $]0, 1[$. Since

$$\mathbb{P}_f(T_\alpha \leq 0) = \mathbb{P}_f\left(\sup_{m \in \mathcal{M}} (\widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} f_0(X_i) - t_m(u_\alpha)) \leq 0\right),$$

we have

$$\mathbb{P}_f(T_\alpha \leq 0) \leq \inf_{m \in \mathcal{M}} \mathbb{P}_f\left(\widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} f_0(X_i) - t_m(u_\alpha) \leq 0\right).$$

Since $\|f - \Pi_{S_m}(f)\|_2^2 = \|f\|_2^2 - \|\Pi_{S_m}(f)\|_2^2$, it holds

$$\begin{aligned}
&\widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^n f_0(X_i) \\
= \;& U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f)) - \|f - \Pi_{S_m}(f)\|_2^2 + \|f\|_2^2 + \|f_0\|_2^2 - 2P_n(f_0) \\
= \;& U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f)) - \|f - \Pi_{S_m}(f)\|_2^2 + \|f - f_0\|_2^2 + 2P(f_0) - 2P_n(f_0),
\end{aligned}$$

which leads to

$$\mathbb{P}_f(T_\alpha \le 0) \le \inf_{m\in\mathcal{M}} \mathbb{P}_f\Bigg(U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f) - 2f) + (P_n - P)(2f - 2f_0) + \|f - f_0\|_2^2 \\ \le \|f - \Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha)\Bigg). \tag{5.29}$$

We then need to control $U_n(H_m)$, $(P_n - P)(2\Pi_{S_m}(f) - 2f)$, $(P_n - P)(2f - 2f_0)$ for every $m \in \mathcal{M}$.

Control of $U_n(H_m)$.
H_m is π-canonical and a direct application of Theorem 5.1 leads to the following Lemma (the proof of Lemma 5.21 is postponed to Section 5.8.2).

Lemma 5.21. *Let us assume that the stationary distribution of the Markov chain $(X_i)_{i\ge1}$ has density f with respect to the Lebesgue measure on $\mathbb{R}$. For all $m = (l, D)$ with $l \in \{1, 2, 3\}$ and $D \in \mathbb{D}_l$, introduce $\{p_l, l \in \mathcal{L}_m\}$ defined as in page 157 and $Z_m = \frac{1}{n(n-1)}\sum_{i\ne j=1}^n H_m(X_i, X_j)$, with $H_m(x, y) = \sum_{l\in\mathcal{L}_m}(p_l(x) - \langle f, p_l\rangle)(p_l(y) - \langle f, p_l\rangle)$. There exist some constants $C, \beta > 0$ (both depending on the Markov chain $(X_i)_{i\ge1}$ while C also depends on ϕ) such that, for all $l \in \{1, 2, 3\}$, $D \in \mathbb{D}_l$ and $u \ge 1$, it holds with probability at least $1 - \beta e^{-u}\log n$,*

$$|Z_{(l,D)}| \le C\left(\|f\|_\infty + 1\right) DR(n, u),$$

where $R(n, u) = \log n\left\{\frac{u}{n} + \left(\frac{u}{n}\right)^2\right\}$.

We deduce that there exist $C, \beta > 0$ such that for any $\gamma \in (0, 1 \wedge (e^{-1}3\beta\log n))$ and any $m = (l, D) \in \mathcal{M}$,

$$\mathbb{P}_f\left(U_n(H_m) \le -C\left(\|f\|_\infty + 1\right) DR\left(n, \log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\right) \le \gamma/3. \tag{5.30}$$

From Eq.(5.29) and Eq.(5.30) we get that

$$\mathbb{P}_f(T_\alpha \le 0) \le \frac{\gamma}{3} + \inf_{m\in\mathcal{M}} \mathbb{P}_f\Bigg((P_n - P)(2\Pi_{S_m}(f) - 2f) + (P_n - P)(2f - 2f_0) + \|f - f_0\|_2^2 \\ \le \|f - \Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha) + C\left(\|f\|_\infty + 1\right) DR\left(n, \log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\Bigg). \tag{5.31}$$

Control of $(P_n - P)(2\Pi_{S_m}(f) - 2f)$.
It is easy to check that there exists some constant $C' > 0$ such that for all l in $\{1, 2\}$, D in $\mathbb{D}_l$,

$$\left|2\Pi_{S_{(l,D)}}(f)(X_i) - 2f(X_i)\right| \le C'\|f\|_\infty.$$

Indeed,

- when $l = 1$, for any $k \in \mathbb{Z}$,

$$\langle\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}, f\rangle = \int \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x)f(x)dx \le D^{-1/2}\|f\|_\infty.$$

Hence,

$$\sup_x |\Pi_{S_{(1,D)}}(f)(x)| \leq \sup_x \sum_{k\in\mathbb{Z}} \left|\langle \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}, f\rangle\right| \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x)$$
$$\leq \quad D^{-1/2}\|f\|_\infty \sup_x \sum_{k\in\mathbb{Z}} \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x) = \|f\|_\infty.$$

- when $l = 2$, $D = 2^J$ for some $J \in \mathbb{N}$ and we have for any $k \in \mathbb{Z}$,

$$\langle \phi_{J,k}, f\rangle = \int 2^{J/2}\phi(2^J x - k) f(x)dx \leq \|f\|_\infty \int 2^{J/2}|\phi(2^J x)|dx \leq 2^{-J/2}\|f\|_\infty \|\phi\|_1.$$

Hence,

$$\sup_x |\Pi_{S_{(2,D)}}(f)(x)| \leq \sup_x \sum_{k\in\mathbb{Z}} |\langle \phi_{J,k}, f\rangle| \times |\phi_{J,k}(x)|$$
$$\leq \quad 2^{-J/2}\|f\|_\infty \|\phi\|_1 \sup_x \sum_{k\in\mathbb{Z}} |2^{J/2}\phi(2^J x - k)| \leq c\|f\|_\infty \|\phi\|_1,$$

where $c > 0$ is a constant depending only on ϕ since ϕ is bounded and compactly supported. Stated otherwise, there is only a finite number of integers $k \in \mathbb{Z}$ (which is independent of x and J) such that for any $x \in \mathbb{R}$ and any $J \in \mathbb{Z}$, $2^J x - k$ falls into the support of ϕ.

Moreover, it is proved in DeVore and Lorentz [1993, Page 269], that one can take C' such that for all D in $\mathbb{D}_3$,

$$|2\Pi_{S_{(3,D)}}(f)(X_i) - 2f(X_i)| \leq C'\|f\|_\infty \log(D+1).$$

Since

$$\mathbb{E}_{X\sim\pi}\left(2\Pi_{S_m}(f)(X) - 2f(X)\right)^2 \leq 4\|f\|_\infty \|\Pi_{S_m}(f) - f\|_2^2,$$

we can deduce using Proposition A.17 (see Section A.5.2) that for all $m = (l, D) \in \mathcal{M}$,

$$\mathbb{P}_f\Bigg((P_n - P)(2\Pi_{S_m}(f) - 2f) < -\frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty \log(D+1)}{n}$$
$$- 2\sqrt{\frac{2\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{n}}\|\Pi_{S_m}(f) - f\|_2\Bigg) \leq \frac{\gamma}{3}.$$

Considering some $\epsilon \in]0, 2[$, we use the inequality $\forall a, b \in \mathbb{R}$, $2ab \leq 4a^2/\epsilon + \epsilon b^2/4$ and we obtain that for any $m = (l, D) \in \mathcal{M}$,

$$\mathbb{P}_f\Bigg((P_n - P)(2\Pi_{S_m}(f) - 2f) + \frac{\epsilon}{4}\|\Pi_{S_m}(f) - f\|_2^2 < -\frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty \log(D+1)}{n}$$
$$- \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\epsilon n}\Bigg) \leq \frac{\gamma}{3}. \tag{5.32}$$

The control of $(P_n - P)(2f - 2f_0)$ is computed in the same way and we get

$$\mathbb{P}_f\Bigg((P_n - P)(2f - 2f_0) + \frac{\epsilon}{4}\|f - f_0\|_2^2 < -\frac{4\log(3C_\chi/\gamma)qA_1(\|f\|_\infty + \|f_0\|_\infty)}{n}$$
$$- \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\epsilon n}\Bigg) \leq \frac{\gamma}{3}. \tag{5.33}$$

Finally, we deduce from Eq.(5.31), Eq.(5.32) and Eq.(5.33) that if there exists some $m = (l, D)$ in $\mathcal{M}$ such

that

$$\left(1-\frac{\epsilon}{4}\right)\|f-f_0\|_2^2 > \left(1+\frac{\epsilon}{4}\right)\|f-\Pi_{S_m}(f)\|_2^2 + \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\epsilon n} + \frac{4\log(3C_\chi/\gamma)qA_1(\|f\|_\infty+\|f_0\|_\infty)}{n} + \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\epsilon n} + \frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty\log(D+1)}{n} + t_m(u_\alpha) + C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right),$$

i.e. such that

$$\left(1-\frac{\epsilon}{4}\right)\|f-f_0\|_2^2 > \left(1+\frac{\epsilon}{4}\right)\|f-\Pi_{S_m}(f)\|_2^2 + \frac{16\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\epsilon n} + 4\left(\|f\|_\infty(C'\log(D+1)+1)+\|f_0\|_\infty\right)\frac{\log(3C_\chi/\gamma)qA_1}{n} + t_m(u_\alpha) + C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right),$$

then $\mathbb{P}_f\left(T_\alpha \leq 0\right) \leq \gamma$.
To conclude the proof of Theorem 5.12, it suffices to notice that for any $\epsilon \in]0,2[$, choosing $\eta > 0$ such that $1+\eta = \frac{1+\frac{\epsilon}{4}}{1-\frac{\epsilon}{4}}$ leads to $\epsilon = \frac{4\eta}{2+\eta}$. One can immediately check that the condition $\epsilon \in]0,2[$ is equivalent to $\eta \in]0,2[$. Noticing further that $\frac{1}{\epsilon} = \frac{2+\eta}{4\eta} < \frac{2+2}{4\eta} = \frac{1}{\eta}$, we deduce that for any $\eta \in]0,2[$, if

$$\|f-f_0\|_2^2 > (1+\eta)\Bigg\{\|f-\Pi_{S_m}(f)\|_2^2 + \frac{16\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\eta n} + 4\left(\|f\|_\infty(C'\log(D+1)+1)+\|f_0\|_\infty\right)\frac{\log(3C_\chi/\gamma)qA_1}{n} + t_m(u_\alpha) + C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\Bigg\},$$

then $\mathbb{P}_f\left(T_\alpha \leq 0\right) \leq \gamma$.

5.8.2 Proof of Lemma 5.21

Lemma 5.21 will follow from Theorem 5.1 if we can show that the function H_m is bounded. Let us denote $m = (l, D)$ for some $l \in \{1,2,3\}$ and $D \in \mathbb{D}_l$. Let us first remark that the Bessel's inequality states that

$$\sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \leq \|f\|_2^2 = \int f(x)f(x)dx \leq \|f\|_\infty, \tag{5.34}$$

since $\int f(x)dx = 1$ and $f(x) \geq 0, \quad \forall x$.
- If $l = 1$, then we notice that for any $k \in \mathbb{Z}$,

$$\begin{aligned}|\langle\sqrt{D}\mathbb{1}_{]k/D,(k+1)/D[}, f\rangle| &= \left|\int \sqrt{D}\mathbb{1}_{]k/D,(k+1)/D[}(x)f(x)dx\right| \\ &\leq \|f\|_\infty\sqrt{D}\int \mathbb{1}_{]k/D,(k+1)/D[}(x)dx \\ &\leq D^{-1/2}\|f\|_\infty.\end{aligned}$$

Then for any $x, y \in \mathbb{R}$ it holds

$$\begin{aligned}|H_m(x,y)| &\leq \sum_{k\in\mathcal{L}_m} |p_k(x)p_k(y)| + \sum_{k\in\mathcal{L}_m} |p_k(x)\langle p_k, f\rangle| + \sum_{k\in\mathcal{L}_m} |p_k(y)\langle p_k, f\rangle| + \sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \\ &\leq \sum_{k\in\mathbb{Z}} D\mathbb{1}_{]k/D,(k+1)/D[}(x)\mathbb{1}_{]k/D,(k+1)/D[}(y) \\ &\quad + 2\sup_z \sum_{k\in\mathbb{Z}} \sqrt{D}|\mathbb{1}_{]k/D,(k+1)/D[}(z)| \times |\langle \sqrt{D}\mathbb{1}_{]k/D,(k+1)/D[}, f\rangle| + \sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \\ &\leq D + 2\|f\|_\infty + \|f\|_\infty,\end{aligned}$$

where in the last inequality we used Eq.(5.34).

- If $l = 2$ then $D = 2^J$ for some $J \in \mathbb{N}$ and we have for any $k \in \mathbb{Z}$,

$$\langle \phi_{J,k}, f\rangle = \int 2^{J/2}\phi(2^J x - k)f(x)dx \leq \|f\|_\infty \int 2^{J/2}|\phi(2^J x)|dx \leq 2^{-J/2}\|f\|_\infty \|\phi\|_1.$$

We get that for any $x, y \in \mathbb{R}$,

$$\begin{aligned}|H_m(x,y)| &\leq \sum_{k\in\mathcal{L}_m} |p_k(x)p_k(y)| + \sum_{k\in\mathcal{L}_m} |p_k(x)\langle p_k, f\rangle| + \sum_{k\in\mathcal{L}_m} |p_k(y)\langle p_k, f\rangle| + \sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \\ &\leq \sum_{k\in\mathbb{Z}} 2^J\phi(2^J x - k)\phi(2^J y - k) + 2\sup_z \sum_{k\in\mathbb{Z}} 2^{-J/2}\|f\|_\infty\|\phi\|_1 2^{J/2}|\phi(2^{J/2}z - k)| \\ &\quad + \sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \\ &\leq c2^J + c'\|\phi\|_1\|f\|_\infty + \|f\|_\infty \\ &= cD + c'\|\phi\|_1\|f\|_\infty + \|f\|_\infty,\end{aligned}$$

for some constants $c, c' > 0$. In the last inequality we used Eq.(5.34) and the fact ϕ is bounded and compactly supported. Indeed, this implies that there is only a finite number of integers $k \in \mathbb{Z}$ (which is independent of x and J) such that for any $x \in \mathbb{R}$ and any $J \in \mathbb{Z}$, $2^J x - k$ falls into the support of ϕ.

- If $l = 3$ then we easily get for any $x, y \in [0, 1]$,

$$\begin{aligned}|H_m(x,y)| &\leq \sum_{k\in\mathcal{L}_m} |p_k(x)p_k(y)| + \sum_{k\in\mathcal{L}_m} |p_k(x)\langle p_k, f\rangle| + \sum_{k\in\mathcal{L}_m} |p_k(y)\langle p_k, f\rangle| + \sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \\ &\leq 2D + 4D\|f\|_\infty + \|f\|_\infty.\end{aligned}$$

We deduce that in any case, H_m is bounded by $c(1 + \|f\|_\infty)D$ for some constant $c > 0$ (depending only on ϕ) which concludes the proof of Lemma 5.21.

5.8.3 Proof of Corollary 5.14

Step 1: We start by providing an upper bound on $t_m(u_\alpha)$ with Lemma 5.22.

Lemma 5.22. *There exists a constant $C(\alpha) > 0$ such that for any $m = (l, D) \in \mathcal{M}$ it holds,*

$$t_m(u_\alpha) \leq W_m(\alpha),$$

where

$$W_m(\alpha) = C(\alpha)\left(\|f_0\|_\infty + 1\right)\left[DR\left(n, \log\log n\right) + \frac{\log\log n}{n}\right].$$

Proof of Lemma 5.22. Let us recall that $t_m(u)$ denotes the $(1-u)$ quantile of the distribution of $\widehat{T}_m$ under

the null hypothesis. One can easily see that $|\mathcal{M}| \leq 3(1+\log_2 n)$. So, setting $\alpha_n = \alpha/(3(1+\log_2 n))$,

$$\begin{aligned}\mathbb{P}_{f_0}(\sup_{m\in\mathcal{M}}(\widehat{T}_m - t_m(\alpha_n)) > 0) &\leq \sum_{m\in\mathcal{M}} \mathbb{P}_{f_0}(\widehat{T}_m - t_m(\alpha_n) > 0)\\ &\leq \sum_{m\in\mathcal{M}} \alpha/(3(1+\log_2 n))\\ &\leq \alpha.\end{aligned}$$

By definition of u_α, this implies that $\alpha_n \leq u_\alpha$ and for all $m \in \mathcal{M}$,

$$t_m(u_\alpha) \leq t_m(\alpha_n).$$

Hence it suffices to upper bound $t_m(\alpha_n)$. Let $m = (l, D) \in \mathcal{M}$. We use the same notation as in the proof of Theorem 5.12 to obtain that

$$\widehat{T}_m = U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f)) - 2P_n(f_0) + \|f_0\|_2^2 + \|\Pi_{S_m}(f)\|_2^2.$$

Under the null hypothesis, this reads as

$$\begin{aligned}\widehat{T}_m &= U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) - \|f_0\|_2^2 + \|\Pi_{S_m}(f_0)\|_2^2\\ &= U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) - \|f_0 - \Pi_{S_m}(f_0)\|_2^2.\end{aligned}$$

We control $U_n(H_m)$ and $(P_n - P)(2\Pi_{S_m}(f_0) - 2f_0)$ exactly like in the proof of Theorem 5.12.
From Lemma 5.21, there exist $C, \beta > 0$ such that for any $m = (l, D) \in \mathcal{M}$, it holds

$$\mathbb{P}_{f_0}\left(U_n(H_m) \leq C\left(\|f_0\|_\infty + 1\right) DR\left(n, \log\left\{\frac{2\beta\log n}{\alpha_n}\right\}\right)\right) \leq \alpha_n/2. \tag{5.35}$$

Moreover, since

$$|2\Pi_{S_{(l,D)}}(f_0)(X_i) - 2f_0(X_i)| \leq C'\|f_0\|_\infty \log(D+1),$$

and

$$\mathbb{E}_{X\sim\pi}\left(2\Pi_{S_m}(f_0)(X) - 2f_0(X)\right)^2 \leq 4\|f_0\|_\infty \|\Pi_{S_m}(f_0) - f_0\|_2^2,$$

we get using Proposition A.17 (see Section A.5.2) that for all $m = (l, D) \in \mathcal{M}$,

$$\begin{aligned}\mathbb{P}_{f_0}\Bigg((P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) > &\frac{2C'\log(2C_\chi/\alpha_n) q A_1 \|f_0\|_\infty \log(D+1)}{n}\\ &+ 2\sqrt{\frac{2\log(2C_\chi/\alpha_n) q A_2 \|f_0\|_\infty}{n}}\|\Pi_{S_m}(f_0) - f_0\|_2\Bigg) \leq \frac{\alpha_n}{2}.\end{aligned}$$

Using the inequality $\forall a, b \in \mathbb{R},\ 2ab \leq a^2 + b^2$, and the fact that for $n \geq 16,\ \log(D+1) \leq \log(n^2+1)$, we obtain that there exists $C'' > 0$ such that

$$\mathbb{P}_{f_0}\left((P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) - \|\Pi_{S_m}(f_0) - f_0\|_2^2 > \frac{C''\|f_0\|_\infty \log(2C_\chi/\alpha_n)\log(n)}{n}\right) \leq \frac{\alpha_n}{2}.$$

We deduce that it holds

$$\mathbb{P}_{f_0}\left(\widehat{T}_m > C\left(\|f_0\|_\infty + 1\right) DR\left(n, \log\left\{\frac{2\beta\log n}{\alpha_n}\right\}\right) + \frac{C''\|f_0\|_\infty \log(2C_\chi/\alpha_n)\log(n)}{n}\right) \leq \alpha_n.$$

Noticing that there exists some constant $c(\alpha) > 0$ such that

$$\log\left\{\frac{2\beta\log n}{\alpha_n}\right\} \vee \log(2C_\chi/\alpha_n) \leq c(\alpha)\log\log n,$$

we deduce by definition of $t_m(\alpha_n)$ that for some $c(\alpha) > 0$,

$$t_m(\alpha_n) \leq c(\alpha)C\left(\|f_0\|_\infty + 1\right) DR\left(n, \log\log n\right) + c(\alpha)\frac{C''\|f_0\|_\infty \log\log n}{n}.$$

□

Step 2: Proof of Corollary 5.14.
Let us fix $\gamma \in]0,1[$ and $l \in \{1,2,3\}$. From Theorem 5.12 and Lemma 5.22, we deduce that if f satisfies

$$\|f - f_0\|_2^2 > (1+\epsilon) \inf_{D\in\mathcal{D}_l} \|f - \Pi_{S_{(l,D)}}(f)\|_2^2 + W_{(l,D)}(\alpha) + V(l,D)(\gamma),$$

then

$$\mathbb{P}_f(T_\alpha \leq 0) \leq \gamma.$$

It is thus a matter of giving an upper bound for

$$\inf_{D\in\mathcal{D}_l} \left\{\|f - \Pi_{S_{(l,D)}}(f)\|_2^2 + W_{(l,D)}(\alpha) + V_{(l,D)}(\gamma)\right\},$$

when f belongs to some specified classes of functions. Recall that

$$\mathcal{B}_s^{(l)}(P,M) = \{f \in L_2(\mathbb{R}) \mid \forall D \in \mathcal{D}_l, \|f - \Pi_{S_{(l,D)}}(f)\|_2^2 \leq P^2 D^{-2s}, \|f\|_\infty \leq M\}.$$

We now assume that f belongs to $\mathcal{B}_s^{(l)}(P,M)$. Since $\|f - \Pi_{S_{(l,D)}}(f)\|_2^2 \leq P^2D^{-2s}$, we only need an upper bound for

$$\inf_{D\in\mathcal{D}_l} \left\{ P^2D^{-2s} + C(\alpha)\left(\|f_0\|_\infty + 1\right)\left[DR\left(n, \log\log n\right) + \frac{\log\log n}{n}\right] + C_1\|f\|_\infty \frac{\log(3C_\chi/\gamma)}{\epsilon n} \right.$$
$$\left. + C_2\left(\|f\|_\infty \log(D+1) + \|f_0\|_\infty\right)\frac{\log(3C_\chi/\gamma)}{n} + C_3\left(\|f\|_\infty + 1\right) DR\left(n, \log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\right\}.$$

Using that f belongs to $\mathcal{B}_s^{(l)}(P,M)$ and the fact that

$$R\left(n, \log\log n\right) \vee R\left(n, \log\left\{\frac{3\beta\log n}{\gamma}\right\}\right) \lesssim \log(n)\frac{\log\log n}{n},$$

where $\lesssim$ states that the inequality holds up to some multiplicative constant independent of n, D and P, we deduce that we want to upper bound

$$\inf_{D\in\mathcal{D}_l} \left\{ P^2D^{-2s} + D\log(n)\frac{\log\log n}{n} + \frac{\log\log n}{n} + \frac{\log(D+1)}{n} \right\}.$$

Since $\log(D+1) \leq D$ for all $D \in \mathcal{D}_l$, we only need to focus on

$$\inf_{D\in\mathcal{D}_l} \left\{ P^2D^{-2s} + D\log(n)\frac{\log\log n}{n} \right\}.$$

Let us point out that $P^2D^{-2s} < D\log(n)\frac{\log\log n}{n}$ if and only if $D > \left(\frac{P^4n^2}{\log^2(n)(\log\log n)^2}\right)^{\frac{1}{4s+2}}$. Hence we define D_* by

$$\log_2(D_*) := \lfloor \log_2\left(\left(\frac{P^4n^2}{\log^2(n)(\log\log n)^2}\right)^{\frac{1}{4s+2}}\right) \rfloor + 1.$$

We consider three cases.

- If $D_* < 1$, then $P^2D^{-2s} < D\log(n)\frac{\log\log n}{n}$ for any $D \in \mathcal{D}_l$ and by choosing $D_0 = 1$ to upper bound

the infimum we get

$$\inf_{D\in\mathcal{D}_l}\left\{\|f-\Pi_{S_{(l,D)}}(f)\|_2^2+W_{(l,D)}(\alpha)+V_{(l,D)}(\gamma)\right\}\leq \log(n)\frac{\log\log n}{n}.$$

- If $D_*>2^{\lfloor \log_2(n/(\log(n)\log\log n)^2)\rfloor}$, then $P^2D^{-2s}>D\log(n)\frac{\log\log n}{n}$ for any $D\in\mathcal{D}_l$ and by choosing $D_0=2^{\log_2(\lfloor n/(\log(n)\log\log n)^2)\rfloor)}$ to upper bound the infimum we get

$$\begin{aligned}\inf_{D\in\mathcal{D}_l}\left\{\|f-\Pi_{S_{(l,D)}}(f)\|_2^2+W_{(l,D)}(\alpha)+V_{(l,D)}(\gamma)\right\}&\lesssim 2P^2D_0^{-2s}\\&\leq 2^{2s+1}P^2\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{2s}.\end{aligned}$$

- Otherwise D_* belongs to $\mathcal{D}_l$ and we upper bound the infimum by choosing $D_0=D_*$ and we get

$$\inf_{D\in\mathcal{D}_l}\left\{\|f-\Pi_{S_{(l,D)}}(f)\|_2^2+W_{(l,D)}(\alpha)+V_{(l,D)}(\gamma)\right\}\lesssim 4P^{\frac{2}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{2s}{2s+1}}.$$

The proof of Corollary 5.14 ends with simple computations that we provide below for the sake of completeness. Since

$$\begin{aligned}&\log(n)\frac{\log\log n}{n}\leq P^{\frac{2}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{2s}{2s+1}}\\\Leftrightarrow&\left(\log(n)\frac{\log\log n}{n}\right)^{1/2}\leq P.\end{aligned}$$

and since

$$\begin{aligned}&P^2\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{2s}&&\leq P^{\frac{2}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{2s}{2s+1}}\\\Leftrightarrow\quad&P\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{s}&&\leq P^{\frac{1}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}}\\\Leftrightarrow\quad&P^{2s}\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{s(2s+1)}&&\leq\left(\frac{\log(n)\log\log n}{n}\right)^{s}\\\Leftrightarrow\quad&P\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{s+1/2}&&\leq\left(\frac{\log(n)\log\log n}{n}\right)^{1/2}\\\Leftrightarrow\quad&P&&\leq\frac{n^s}{(\log(n)\log\log n)^{2s+1/2}},\end{aligned}$$

we deduce that if P is chosen such that

$$\left(\log(n)\frac{\log\log n}{n}\right)^{1/2}\leq P\leq\frac{n^s}{(\log(n)\log\log n)^{2s+1/2}},\tag{5.36}$$

then the uniform separation rate of the test $\mathbb{1}_{T_\alpha>0}$ over $\mathcal{B}_s^{(l)}(P,M)$ satisfies

$$\rho\left(\mathbb{1}_{T_\alpha>0},\mathcal{B}_s^{(l)}(P,M),\gamma\right)\leq C'P^{\frac{1}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}}.\tag{5.37}$$

Remark. This final statement can allow the reader to understand our choice for the size of the model $|\mathcal{M}|$ that we considered. Indeed, we chose for any $l\in\{1,2,3\}$, $\mathcal{D}_l=\{2^J,0\leq J\leq\log_2\left(n/(\log(n)\log\log n)^2\right)\}$ in order to ensure that for values of P saturing the right inequality in (5.36) (i.e. for $P\approx\frac{n^s}{(\log(n)\log\log n)^{2s+1/2}}$), the upper-bound in Eq.(5.37) still tends to zero as n goes to $+\infty$ for any possible values of the smoothness parameter s.

Chapter 6

Selective Inference with the Generalized Linear Lasso

Chapter Abstract

We investigate the distribution of the solutions of the generalized linear lasso (GLL), conditional on some selection event. In this framework of post-selection inference (PSI), we provide rigorous definitions of the selected and saturated models: two different paradigms that determine the hypothesis being tested. Based on a conditional Maximum Likelihood Estimator (MLE) approach, we give a procedure to obtain asymptotically valid PSI confidence regions and testing procedures for Generalized Linear Models (GLMs). In a second stage, we focus on the sparse logistic regression and we exhibit conditions ensuring that our conditional MLE method is valid. We present numerical simulations supporting our theoretical results.

Chapter Content

6.1 Friendly introduction to post-selection inference

This chapter was originally motivated by the problem of error quantification for link prediction in random graphs as explained in Section 1.3. It appears that the questions we asked ourselves in the context of random graphs can be tackled in a more general framework, namely post-selection inference in generalized linear models. In this chapter, we will keep this level of generality and we encourage the reader to read Section 1.3 to understand how our contributions find applications in the context of networks.

In this first section, we aim at providing a gentle introduction to the field of selective inference. The section consists of three steps from a concrete motivation for post-selection inference to a mathematical analysis of a specific framework. First, we describe the problem of publication bias in the scientific literature that highlights the importance of selective inference. Next, we present the general principles of selective inference. In a final subsection, we present exact methods for post-selective inference in the linear Gaussian model.

6.1.1 The file drawer effect

In medicine, physics or engineering, one often needs to measure some quantity in order to detect a suspected effect or gain information about a known one. Due to perturbations errors and noise coming from instrument or environment, it is necessary to repeat several times the experiment in order to average the fluctuations. It appears that it is not always possible (or desirable) for a single research team to conduct enough simulations to draw a faithful conclusion on the presence or absence of the studied effect. This can be typically due to financial or time reasons. In this context, the standard approach consists in combining the results of different measurements of the same effect using the data collected from the literature in order to improve the signal-to-noise ratio. The problem with this approach is that the results collected from the literature are often not a representative sample. The reason is that works presenting results that are not statistically significant are more likely to be unpublished (cf. Scargle [2000]).
Let us recall that for a significance level of 5%, then in repeated studies, about 5% of studies of a situation where the null hypothesis is true will falsely reject the null hypothesis. Thus, if just (or even predominantly) the statistically significant studies are published, the published record mis-represents the true situation. This phenomenon - called the *publication bias* in the literature - can have catastrophic scientific consequences. In particular,

- effects that are not real may appear to be supported by research,
- research teams could put a lot of effort answering a question that has been already studied but not reported.

To cope with this issue, several journals are now completely devoted to the publication of studies with negative results, among them, *The Missing Pieces: A Collection of Negative, Null and Inconclusive Results* and *The All Results Journal*.
Let us now tackle the publication bias with a more mathematical perspective. Note that this discussion is mainly inspired by [Tian and Taylor, 2015, Example 1]. We consider n different scientific research groups and we assume that each of them makes m independent measurements of some quantity of interest, whose true value is $\mu \in \mathbb{R}$. We consider that the measurements of i-th team are given by $(Y_k^{(i)})_{k\in[m]}$ and we assume that $Y_k^{(i)} \sim \mathcal{N}(\mu, 1)$. Due to the publication bias, only the teams reporting measurements whose sample mean $\overline{Y}^{(i)} = \frac{1}{m}\sum_{k=1}^m Y_k^{(i)}$ survives the file drawer effect

$$\{i \in [n] \,|\, |\sqrt{m}\,\overline{Y}^{(i)}| > 1\},$$

get their paper published. Let us recall that we aim at testing with level $\alpha = 5\%$ the null $\mu = 0$ (absence of effect) against the alternative $\mu \neq 0$ using the data collected from the published papers. In this context, what is the correct rejection region allowing to obtain a testing procedure correctly calibrated? To solve this question, we want to compute the positive real $c > 0$ such that

$$\mathbb{P}_{\mu=0}(|\sqrt{m}\,\overline{Y}^{(1)}| > c \,|\, |\sqrt{m}\,\overline{Y}^{(1)}| > 1) = 0.05.$$

Table 6.1 shows the difference in the rejection regions to obtain a well-calibrated test with or without publication bias.

	c
$\mathbb{P}(\lvert\sqrt{m}\,\overline{Y}\rvert > c \mid \lvert\sqrt{m}\,\overline{Y}\rvert > 1) = 5\%$	$2,43$
$\mathbb{P}(\lvert\sqrt{m}\,\overline{Y}\rvert > c) = 5\%$	$1,96$

Table 6.1: Threshold c for the rejection region $|\sqrt{m}\,\overline{Y}| > c$ ensuring a Type I error of 5% when the distribution of $\sqrt{m}\,\overline{Y}$ is either $\mathcal{N}(0,1)$ or $\mathcal{N}(0,1)$ conditional on the event $\{|\sqrt{m}\,\overline{Y}| > 1\}$.

The example of the file drawer problem shed light on the necessity to take into account the potential selection procedure to provide valid inference methods.

6.1.2 Post-selection inference for statisticians

The old fashion to conduct inference was to collect data and then to test hypotheses. However as time goes along, it becomes always cheaper to store large amount of data and statisticians are more and more coping with high-dimensional problems, making most of standard inferences produced by common software packages often unreliable [cf. Sur and Candès, 2019]. For this reason, it is now common to perform some model selection step before making inference. This dimension reduction method allows to circumvent the curse of dimensionality and to compute estimates relatively efficiently. Of course, the model selection step is performed by looking at the data. Considering the example of maximum likelihood estimation with a sparsity inducing norm, one needs to choose the hyper-parameter that drives the tradeoff between the data-fidelity term and the regularization. This choice will lead to a set of *selected variables*: the ones with non-zero coefficients for the optimal solution of the penalized likelihood optimization problem.
A standard approach to provide valid inference procedure in this context is to use the so called *data splitting method*. The idea is simply to split the data into two parts. The first part will be used to select the model while the second part will be used for inference using the selected model with classical statistical tools. This procedure is widely used in practice despite the paucity of the literature on the subject. As explained in Fithian et al. [2014], data splitting solves the problem of controlling selective errors but at a cost: we are tucking away a significant part of the data in the inference step (the one we used to select the model). But this remark is also true at the first stage of the procedure (!) where we do not use the data reserved for inference in the model selection step. Last but not least, some data are structured in such a way that is not possible to split it into independent parts. This is the case for example for time series.
For these reasons, post-selection inference aims at using all the data for both the model selection and the inference steps. In this context, one needs to account for the model selection step to propose correctly calibrated inference procedures. Typically, in the inference stage, we will condition the distribution of the observed response on the so-called *selection event*. The selection event is defined as the set of observations that would have led to the same selected model. Table 6.2 sheds light on the different use of the data for PSI or for data splitting. In the next paragraph, we propose a brief presentation of exact post-selection inference in the linear model with Gaussian noise.

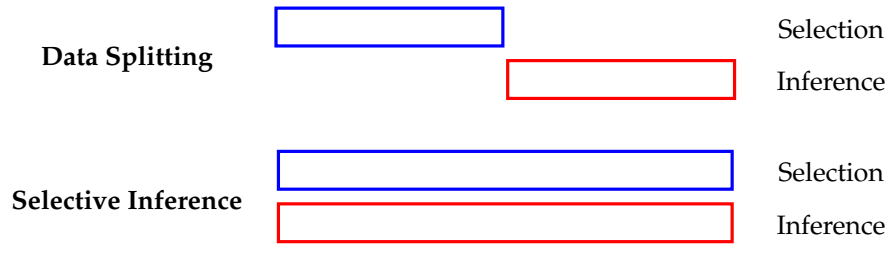


Table 6.2: Usage of the data for the selection and inference stages for data splitting and post-selection inference.

6.1.3 Post-selection inference for the sparse linear regression with Gaussian noise

Gaussian linear model: from selection to estimation. In linear regression with Gaussian noise, the data arise from a multivariate normal distribution

$$Y \sim \mathcal{N}(\mathbf{X}\vartheta^*, \sigma^2 \mathrm{Id}_N),$$

where $\mathbf{X} \in \mathbb{R}^{N\times d}$ is the design matrix. In this succinct presentation, we will assume that $\sigma^2 > 0$ is known (and we refer to [Fithian et al., 2014, Section 4] for the interested reader that wants to tackle the case where σ^2 is unknown). We further consider that the columns of $\mathbf{X}$ are in general positions [cf. Tibshirani, 2013, Section 2.2], meaning that the affine span of any $k+1$ points $\sigma_1\mathbf{X}_{i_1}, \ldots, \sigma_{k+1}\mathbf{X}_{i_{k+1}}$, for arbitrary signs $\sigma_1, \ldots, \sigma_{k+1} \in \{-1, 1\}$, does not contain any element of $\{\pm\mathbf{X}_i \,:\, i \notin \{i_1, \ldots, i_{k+1}\}\}$. Equivalently, this assumption means that no k-dimensional subspace $L \subset \mathbb{R}^N$, for $k < \min(N, d)$, contains more than $k+1$ elements of the set $\{\pm\mathbf{X}_1, \ldots, \pm\mathbf{X}_d\}$, excluding antipodal pairs. Let us point out that the entries of the design matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$ are sampled from a continuous probability distribution on $\mathbb{R}^{N\times d}$, then the columns of $\mathbf{X}$ are in general positions with probability one (see [Tibshirani, 2013, Section 2.2]).
We assume that the analyst selects the model using a ℓ_1 penalty. Namely, given some hyper-parameter $\lambda > 0$, she computes

$$\widehat{\vartheta}^\lambda \in \arg\min_{\vartheta\in\mathbb{R}^d} \left\{ \frac{1}{2}\|Y - \mathbf{X}\vartheta\|_2^2 + \lambda\|\vartheta\|_1 \right\}, \tag{6.1}$$

and defines the set of active variables by

$$M := \widehat{M}(Y) := \{k \in [d] \,|\, \widehat{\vartheta}^\lambda_k \neq 0\}.$$

Once the set of active variables is obtained, the analyst computes the MLE estimate in the linear model with design matrix $\mathbf{X}_M$ where we consider only the features indexes by M, namely denoting $s = |M|$,

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^s} \left\{ \|Y - \mathbf{X}_M\theta\|_2^2 \right\}.$$

Since we assumed that the columns of the design matrix $\mathbf{X}$ are in general positions, we know that $\mathbf{X}_M$ has full column rank [cf. Tibshirani, 2013, Section 2.2] so that the MLE is unique and can be written as

$$\widehat{\theta} = \mathbf{X}_M^+ Y,$$

where $\mathbf{X}_M^+ = \left(\mathbf{X}_M^\top \mathbf{X}_M\right)^{-1}\mathbf{X}_M^\top$ is the pseudo-inverse of the matrix $\mathbf{X}_M$.

Selective inference. In order to provide exact PSI, our goal is first to understand the distribution of some transformation of the MLE $\widehat{\theta}$ conditional on the so-called *selection event* E_M which is defined by

$$E_M := \{Y \in \mathbb{R}^N \,|\, \widehat{M}(Y) = M\}.$$

The selection event E_M corresponds to the set of observations that would have led to the same set of active variables M. Lee et al. [2016] proved that the distribution of a linear transformation of $\widehat{\theta}$ conditional on E_M is known and has a closed-form expression. The next paragraph is dedicated to a short presentation of this result. This allows to conduct selective inference on a linear transformation of $\mathbb{E}[\widehat{\theta}] = \mathbf{X}_M^+\mu^*$, where $\mu^* = \mathbf{X}\vartheta^*$. It becomes crucial at this point to understand the nature of this quantity for which we will provide inference procedures. There are two distinct modeling frameworks [cf. Fithian et al., 2014, Section 4].

- Either we are confident that the selected model is "relevant" in the sense that there exists some $\theta^* \in \mathbb{R}^s$ (and then such vector is unique) such that $\mu^* = \mathbf{X}\vartheta^* = \mathbf{X}_M\theta^*$. In this case, we say that we work under the *selected model* and then we will conduct inference procedure on linear transformations of θ^*. Let us point out that regression coefficients $\theta^*_j = e_j^\top \mathbf{X}_M^+\mu^*$, $j \in [s]$ can be written in this form (where $e_j \in \mathbb{R}^s$ is the vector with all entries set to 0 except the j-th which is 1).
- Or, we do not take the selected model M too seriously and we do not make the assumption that

μ^* belongs to the span of the columns of $\mathbf{X}_M$. In this case, we make inference on $\mathbf{X}_M^+\mu^*$ which simply corresponds to best linear predictor in the population for design matrix $\mathbf{X}_M$. We say in this case that we work under the *saturated model*.

Description of the selection event as a union of polytopes. Since the optimization problem Eq.(6.1) is convex and unconstrained, the Karush-Kuhn and Tucker (KKT) conditions are necessary and sufficient conditions for optimality. One thus obtain that some vector $\widehat{\vartheta}$ is an optimal solution to Eq.(6.1) if and only if there exists $\widehat{S} \in [-1,1]^d$ such that

$$\begin{cases} \mathbf{X}^\top\left(Y - \mathbf{X}\widehat{\vartheta}\right) = \lambda \widehat{S} & \text{(6.2a)} \\ \widehat{S}_k = \operatorname{sign}(\widehat{\vartheta}_k) \qquad \text{if } \widehat{\vartheta}_k \neq 0 & \text{(6.2b)} \\ \widehat{S}_k \in [-1,1] \qquad \text{if } \widehat{\vartheta}_k = 0 & \text{(6.2c)} \end{cases}$$

Since the columns of the design matrix are assumed to be in general positions, we know that the solution to Eq.(6.1) is unique [cf. Tibshirani, 2013, Lemma 3]. From the first KKT condition (see Eq.(6.2a)), we deduce that for any $Y \in \mathbb{R}^N$, there exists a unique vector $\widehat{S} \in [-1,1]^d$ satisfying Eq.(6.2) and we denote this vector $\widehat{S}(Y)$. In the following, we will identify the equicorrelation set defined by

$$\left\{k \in [d] \mid |\widehat{S}_k(Y)| = 1\right\},$$

and the set of predictors with nonzero coefficients $\widehat{M}(Y) = \{k \in [d] \mid \hat{\vartheta}_k^\lambda \neq 0\}$, also called "selected" model. Since $|\widehat{S}_k(Y)| = 1$ for any $\hat{\vartheta}_k^\lambda \neq 0$, the equicorrelation set does in fact contain all predictors with nonzero coefficients, although it may also include some predictors with zero coefficients. However, for almost every λ, the equicorrelation set is precisely the set of predictors with nonzero coefficients, see Lee et al. [2016]. This identification allows us to replace Eq.(6.2c) by

$$\widehat{S}_k \in]-1,1[\quad \text{if } \widehat{\vartheta}_k = 0.$$

By working on the selection event E_M, the KKT conditions can be equivalently written as

$$\begin{cases} \mathbf{X}_M^\top\left(Y - \mathbf{X}_M\widehat{\vartheta}_M\right) = \lambda \widehat{S}_M & \text{(6.3a)} \\ \widehat{S}_M = \operatorname{sign}(\widehat{\vartheta}_M) & \text{(6.3b)} \\ \|\mathbf{X}_{-M}^\top\left(Y - \mathbf{X}_M\widehat{\vartheta}_M\right)\|_\infty < \lambda & \text{(6.3c)} \end{cases}$$

Taking a closer look at the KKT conditions, one can notice that it will be much more convenient to condition on both the selected model E_M and on the vector of signs $\widehat{S}_M$. For this purpose, we denote for any $M \subset [d]$ with cardinality s and any $S_M \in \{-1,1\}^s$,

$$E_M^{S_M} := \{Y \in \mathbb{R}^N \mid \widehat{M}(Y) = M,\ \operatorname{sign}(\widehat{\vartheta})_M = S_M\}.$$

Note that E_M can be recovered from the sets $E_M^{S_M}$, $S_M \in \{\pm 1\}^s$ since

$$E_M = \cup_{S_M \in \{-1,1\}^s} E_M^{S_M}. \tag{6.4}$$

The impressive result obtained from Lee et al. [2016] and known as the *polyhedral lemma* states that the set $E_M^{S_M}$ is a polytope, namely

$$E_M^{S_M} = \{Y \in \mathbb{R}^N \mid A(M,S_M)Y \leq b(M,S_M)\}, \tag{6.5}$$

with

$$A(M,S_M) := \begin{bmatrix} \frac{1}{\lambda}\mathbf{X}_{-M}^\top(\mathrm{Id} - \mathrm{Proj}_{\mathbf{X}_M}) \\ -\frac{1}{\lambda}\mathbf{X}_{-M}^\top(\mathrm{Id} - \mathrm{Proj}_{\mathbf{X}_M}) \\ -\mathrm{Diag}(S_M)\left(\mathbf{X}_M^\top\mathbf{X}_M\right)^{-1}\mathbf{X}_M^\top \end{bmatrix},$$

$$b(M, S_M) := \begin{bmatrix} \mathbf{1} - \mathbf{X}_{-M}^\top \left(\mathbf{X}_M^\top\right)^+ S_M \\ \mathbf{1} + \mathbf{X}_{-M}^\top \left(\mathbf{X}_M^\top\right)^+ S_M \\ -\lambda \mathrm{Diag}(S_M) \left(\mathbf{X}_M^\top \mathbf{X}_M\right)^{-1} S_M \end{bmatrix},$$

and where we have denoted by $\mathrm{Proj}_{\mathbf{X}_M}$ the orthogonal projection on the span of the column of $\mathbf{X}_M$. Hence, the distribution of Y conditionally on $E_M^{S_M}$ is a truncated multivariate normal distribution, truncated to a polytope. If we did not condition on the signs, we would get a multivariate normal truncated to a union of many polytopes, which would not be practical to work with.

Exact PSI methods in the Gaussian linear model. We wish to do inference about some linear transformation of $\mathbf{X}_M^+ \mu^*$ as previously explained. In the selected model, we consider that there exists some $\theta^* \in \mathbb{R}^s$ such that $\mu^* = \mathbf{X}\vartheta^* = \mathbf{X}_M \theta^*$ and one can typically consider $\eta = (\mathbf{X}_M^+)^\top e_j \in \mathbb{R}^N$ (where e_j is the vector with all entries set to 0 except the j-th entry that is set to 1) so that $\eta^\top \mu^* = \theta_j^*$.
A natural statistic to use is $\eta^\top Y$ which is – unconditionally – distributed as $\mathcal{N}(\eta^\top \mu^*, \sigma^2 \|\eta\|_2^2)$. In order to do selective inference, we are interested in the distribution of $\eta^\top Y$ conditionally on $E_M^{S_M}$, which is a complicated mixture of truncated normals that will be computationally expensive to sample from. To make this approach computationally tractable, we also condition on the value of the projection of Y onto the space orthogonal to η, namely $\mathrm{Proj}_\eta^\perp Y$. In the following, we denote by $\mathrm{TN}(a, b, I)$ the distribution of the normal distribution $\mathcal{N}(a, b)$ truncated to the set $I \subset \mathbb{R}$ and by $F_{a,b}^I$ the cumulative distribution function of $\mathrm{TN}(a, b, I)$. The polyhedral lemma can be used to prove that

$$\eta^\top Y \,|\, \{E_M^{S_M}, \mathrm{Proj}_\eta^\perp Y\} \overset{(d)}{=} \mathrm{TN}(\eta^\top \mu^*, \sigma^2 \|\eta\|_2^2, [\mathcal{V}^-(\mathrm{Proj}_\eta^\perp Y, M, S_M), \mathcal{V}^+(\mathrm{Proj}_\eta^\perp Y, M, S_M)]), \tag{6.6}$$

which is a truncated normal distribution for some truncation interval

$$[\mathcal{V}^-(\mathrm{Proj}_\eta^\perp Y, M, S_M), \mathcal{V}^+(\mathrm{Proj}_\eta^\perp Y, M, S_M)].$$

This truncation interval is the line segment which is the intersection of the polytope $\{A(M, S_M)Y \leq b(M, S_M)\}$ with the line spanned by $\mathrm{Proj}_\eta^\perp Y$. Let us stress that this equality in distribution in Eq.(6.6) is not trivial and crucially uses the fact that $\eta^\top Y$ and $\mathrm{Proj}_\eta^\perp Y$ are independent since they are projections of a Gaussian vector with independent components along orthogonal directions. We can then use the probability integral transform to obtain

$$\left[F_{\eta^\top \mu^*, \sigma^2 \|\eta\|_2^2}^{[\mathcal{V}^-(z,M,S_M), \mathcal{V}^+(z,M,S_M)]}(\eta^\top Y) \,|\, \{E_M^{S_M},\ z = \mathrm{Proj}_\eta^\perp(Y)\} \right] \sim \mathcal{U}([0, 1]). \tag{6.7}$$

To reduce the notational clutter, let us denote $F^z(\eta^\top Y) \equiv F_{\eta^\top \mu^*, \sigma^2 \|\eta\|_2^2}^{[\mathcal{V}^-(z,M,S_M), \mathcal{V}^+(z,M,S_M)]}(\eta^\top Y)$ and $Z = \mathrm{Proj}_\eta^\perp Y$. Denoting p_X the density of a random variable X conditional on $E_M^{S_M}$, using Eq.(6.7) we have shown that for any $z \in \mathbb{R}^N$ and any $t \in \mathbb{R}$,

$$p_{F^Z(\eta^\top Y) \,|\, Z=z}(t) = \frac{p_{(F^Z(\eta^\top Y), Z)}(t, z)}{p_Z(z)} = \mathbb{1}_{[0,1]}(t).$$

By integrating over z, we get that

$$p_{F^Z(\eta^\top Y)}(t) = \int_z p_{F^Z(\eta^\top Y) \,|\, Z=z}(t) p_Z(z) dz = \int_z \mathbb{1}_{[0,1]}(t) p_Z(z) dz = \mathbb{1}_{[0,1]}(t).$$

Hence, we proved that

$$\left[F_{\eta^\top \mu^*, \sigma^2 \|\eta\|_2^2}^{[\mathcal{V}^-(Z,M,S_M), \mathcal{V}^+(Z,M,S_M)]}(\eta^\top Y) \,|\, E_M^{S_M} \right] \sim \mathcal{U}([0, 1]). \tag{6.8}$$

Eq.(6.8) shows that we obtained a pivotal quantity that can be used to provide exact post-selection hypothesis testing methods or conditional confidence intervals on $\eta^\top \mu^*$.

What about PSI conditional on E_M? The cautious reader may be troubled by the proposed selective inference methods from the previous paragraph. Indeed, the latter procedures are proved to be correctly calibrated conditional on $E_M^{S_M}$, which is not the selection event E_M. Actually, it is straightforward

to notice that *"inferences that are valid conditional on this finer event will also be valid conditional on* E_M*"* [see Lee et al., 2016, Section 5]. We discuss with further details PSI methods conditional on the signs in Section 6.9. If we want to only condition on the model E_M, then we will have to understand the distribution of $\eta^\top Y$, conditional on Y falling into a union of such polyhedra, that is

$$\eta^\top Y \,|\, E_M \overset{(d)}{=} \eta^\top Y \,|\, \cup_{S_M \in \{-1,+1\}^s} \{A(M, S_M)Y \leq b(M, S_M)\}. \tag{6.9}$$

Using Eq.(6.4) and following a proof analogous to the case where we condition on $E_M^{S_M}$, one can show that it holds

$$\left[F^{\cup_{S_M \in \{\pm 1\}^s}[\mathcal{V}^-(Z,M,S_M), \mathcal{V}^+(Z,M,S_M)]}_{\eta^\top \mu^*, \sigma^2 \|\eta\|_2^2} (\eta^\top Y) \,|\, E_M \right] \sim \mathcal{U}([0,1]), \tag{6.10}$$

where we recall that $|M| = s$.
To sum up the main results of this section, we end up with two different PSI procedures in the Gaussian linear model.

- In case $s = |M|$ is relatively small, one can afford to cover the 2^s vectors of signs $S_M \in \{-1, +1\}^s$ to compute the truncation interval $\cup_{S_M \in \{\pm 1\}^s}[\mathcal{V}^-(Z, M, S_M), \mathcal{V}^+(Z, M, S_M)]$ arising in the pivotal statistic from Eq.(6.10). By inverting this quantity, we obtain intervals with prescribed coverage conditional on E_M.
- To reduce the computational burden of the previous approach, one can simply decide to use the pivotal quantity from Eq.(6.8) to provide valid PSI inference conditional on $E_M^{S_M}$ (and thus also valid conditional on E_M) where S_M is the observed vector of signs. Note that this gain in computational efficiency may be paid in statistical efficiency. When the signal is strong, there will be in general only a small difference in the widths of the conditional confidence intervals obtained using Eq.(6.8) or Eq.(6.10). On the other hand, when the signal is weak, the conditional confidence intervals obtained from Eq.(6.8) will be in general much wider.

Visualization on a specific example. We consider a design matrix $\mathbf{X} \in \mathbb{R}^{2\times 4}$ and we set $\vartheta^* = [0, 0, 0, 0]$. We assume that we observe the vector $Y = [0.2, -0.2]$. Using some regularization parameter $\lambda > 0$, we obtain the following set of active variables $M = \{1, 2\}$ and the observed vector of signs is $\widehat{S}_M(Y) = (-1, +1)$. Using both Eq.(6.4) and Eq.(6.5), we obtain the selection event E_M as the union of polytopes. Each polytope corresponds to a specific vector of signs $S_M \in \{-1, +1\}^2$ as presented in Figure 6.1.

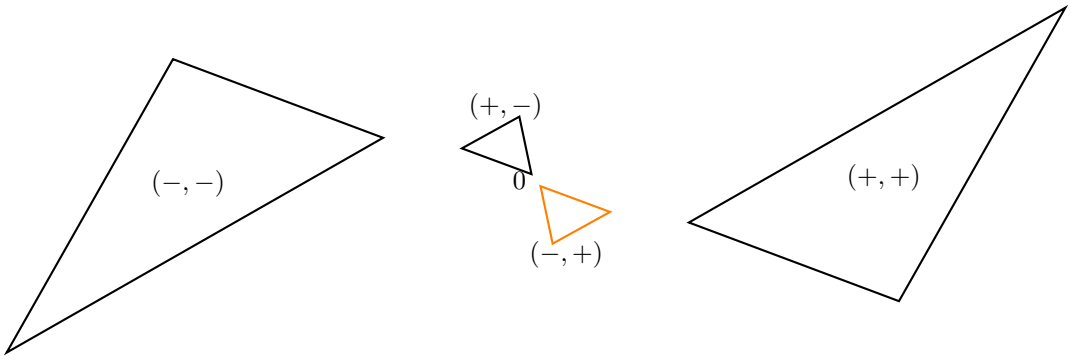


Figure 6.1: A geometric visualization of the selection event E_M. E_M is the union of polytopes (here triangles). Each polytope corresponds to a specific vector of signs $S_M \in \{-1, +1\}^2$.

Since $\vartheta^* = [0, 0, 0, 0]$, we can work under the selected model since $\theta^* = [0, 0]$ always satisfies $\mathbf{X}\vartheta^* = \mathbf{X}_M\theta^*$. We aim at making inference about θ_1^*. To do so, we consider $\eta = (\mathbf{X}_M^+)^\top e_1$. As previously explained, we have two different approaches to make selective inference on θ_1^*.

- Either we decide to design our inference procedure working with the distribution of Y conditional on $E_M^{(-1,+1)}$ where $(-1, +1)$ is the observed vector of signs. We know that the distribution of Y conditional on $E_M^{(-1,+1)}$ is a multivariate Gaussian truncated to the triangle labeled $(-, +)$ on Figure 6.1. Figure 6.2 illustrates the truncated normal distribution from Eq.(6.6) of $\eta^\top Y \,|\, \{E_M^{S_M}, \text{Proj}_\eta^\perp Y\}$.

- Or, we decide to pay an additional computational cost (tiny in this toy example where $s = 2$ is small) to reach higher statistical efficiency, *i.e.* more powerful testing methods or narrower confidence intervals. To do so, we consider the distribution Y conditional on E_M. We know that Y is a mixture of truncated multivariate Gaussians. Each truncated multivariate Gaussian is supported on one of the triangles from Figure 6.1. Figure 6.3 illustrates the distribution of $\eta^\top Y \,|\, \{E_M, \mathrm{Proj}_\eta^\perp Y\}$ (cf. Eq.(6.9)) which is a truncated Gaussian where the truncation set is a union of disjoint intervals.

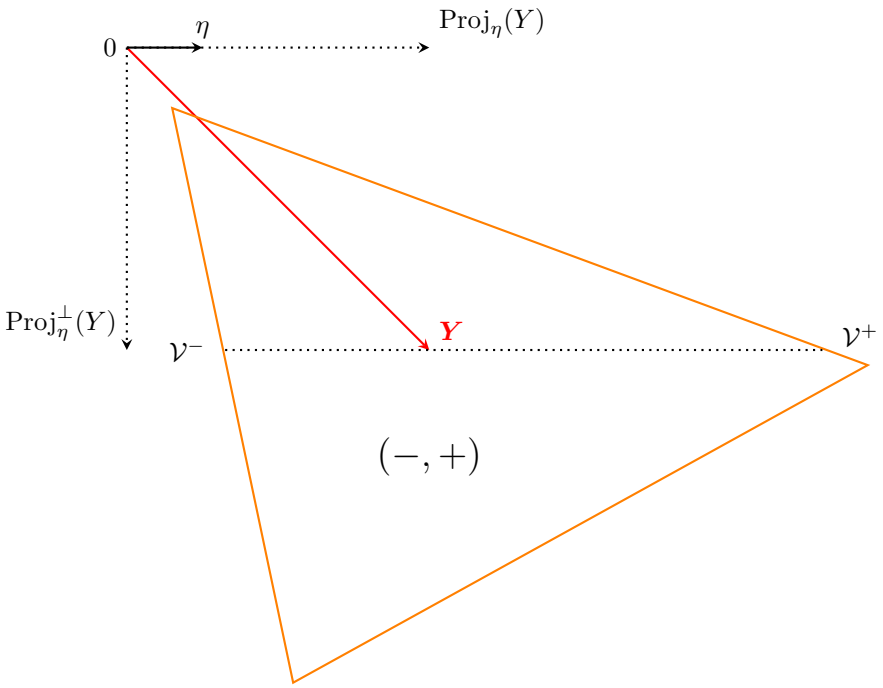


Figure 6.2: A geometric interpretation of why the event $E_M^{(-1,+1)}$ can be characterized as $\{\mathcal{V}^- \leq \eta^\top Y \leq \mathcal{V}^+\}$. $\mathcal{V}^-$ and $\mathcal{V}^+$ are functions of only $\mathrm{Proj}_\eta^\perp(Y)$, which is independent of $\eta^\top Y$.

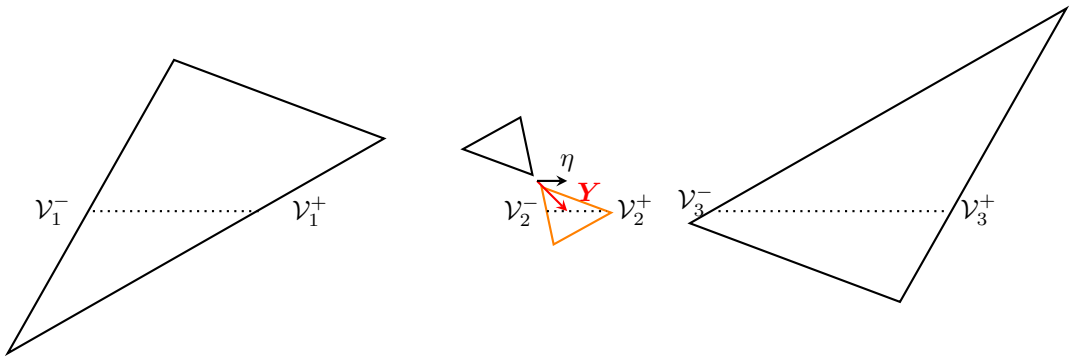


Figure 6.3: When we take the union over signs, the conditional distribution of $\eta^\top Y$ is truncated to a union of disjoint intervals. In this case, the Gaussian is truncated to the set $\cup_{i=1}^3 [\mathcal{V}_i^-, \mathcal{V}_i^+]$.

The previous computations rely heavily on the fact that the model is linear, the noise is assumed to be Gaussian and that the model has been selected using the LASSO. Tackling the problem of post-selection inference outside of this specific setting is much more involved and one should not expect to get exact PSI procedures for an arbitrary generalized linear model (GLM). In this chapter, we aim at pushing the current state of knowledge further regarding PSI methods in GLMs. Let us stress that this chapter is

not disconnected from the rest of this thesis since this work was originally motivated by the problem of link prediction in random graphs. We refer to Section 1.3 for a presentation of the thought process that led us from the link prediction problem on graphs to the results presented in this chapter.

6.2 Introduction

In modern statistics, the number of predictors can far exceed the number of observations available. However, in this high-dimensional context, ℓ_1 regularisation allows for a small number of predictors to be selected (referred to as the selected support) while allowing for a minimax optimal prediction error, see for instance [Van de Geer, 2016, Chapter 2]. The estimated parameters and support are not explicitly known and are obtained by solving a convex optimisation program in practice. This makes inference of the model parameters difficult if not impossible. One solution is to infer conditionally on the selected support. In this framework, it is possible to give a confidence interval and to test any linear statistic.
The ubiquity of the logistic model to solve practical regression problems and the surge of high dimensional data-sets make the sparse logistic regression (SLR) more and more attractive. In this context, it becomes essential to provide certifiable guarantees on the output of the SLR, e.g. confidence intervals.

6.2.1 Post-Selection Inference for high-dimensional generalized linear model

We are interested in a target parameter $\vartheta^\star \in \mathbb{R}^d$ attached to the distribution $\mathbb{P}_{\vartheta^\star}$ of N independent response variables $Y := (y_1, \dots, y_N) \in \mathcal{Y}^N \subseteq \mathbb{R}^N$ given by the data $Z := (z_1, \dots, z_N)$ where $z_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ with $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ a covariate, namely a vector of d predictors. The family of generalized linear models, or GLMs for short, is based on modeling the conditional distribution of the responses $y_i \in \mathcal{Y}$ given the covariate $\mathbf{x}_i \in \mathcal{X}$ in an exponential family form, namely

$$\mathbb{P}_{\vartheta^\star}(y|\mathbf{x}) = h_{\vartheta^\star}(y) \exp\left\{ \frac{y\langle \mathbf{x}, \vartheta^\star \rangle - \xi(\langle \mathbf{x}, \vartheta^\star \rangle)}{v} \right\},$$

where $v > 0$ is a scale parameter, and $\xi : \mathbb{R} \to \mathbb{R}$ is the partition function which is assumed to be of class $\mathcal{C}^{m+1}$ (with m a non-negative integer). With a slight abuse of notations, we will simply denote $\mathbb{P}_{\vartheta^*}(\cdot \,|\, \mathbf{x})$ by $\mathbb{P}_{\vartheta^*}(\cdot)$. Standard examples are $\xi(t) = t^2/2$ for the Gaussian linear model with noise variance v and observation space $\mathcal{Y} = \mathbb{R}$, or $v = 1$, $\xi(t) = \log(1 + \exp(t))$ and $\mathcal{Y} = \{0, 1\}$ for the logistic regression. The negative log-likelihood takes the form

$$\forall \vartheta \in \mathbb{R}^d\,,\ \mathcal{L}_N(\vartheta, Z) := \sum_{i=1}^{N} \xi(\langle \mathbf{x}_i, \vartheta \rangle) - \langle y_i \mathbf{x}_i, \vartheta \rangle\,. \tag{6.11}$$

We assume that the partition function ξ is differentiable, then the score function $\nabla_\vartheta \mathcal{L}_N(\vartheta)$ is given by

$$\forall \vartheta \in \mathbb{R}^d\,,\ \nabla_\vartheta \mathcal{L}_N(\vartheta, Z) = \mathbf{X}^\top \big(\sigma(\mathbf{X}\vartheta) - Y\big)\,,$$

where $\sigma = \xi'$ is the derivative of the partition function and $\mathbf{X} \in \mathbb{R}^{N \times d}$ is referred to as the design matrix whose rows are the covariates and the columns are the predictors. Note that $\sigma(\mathbf{X}\vartheta)$ should be understood as applying entrywise the function σ to the vector $\mathbf{X}\vartheta$. In a high-dimensional context one has more predictors than observations (*i.e.*, $N \ll d$), and one would like to select a small number of predictors to explain the response. We use an ℓ_1-regularization to enforce a structure of sparsity in ϑ. Our overall estimator is based on solving the generalized linear Lasso

$$\hat{\vartheta}^\lambda \in \arg\min_{\vartheta \in \mathbb{R}^d} \big\{ \mathcal{L}_N(\vartheta, Z) + \lambda \|\vartheta\|_1 \big\} \tag{6.12}$$

where $\lambda > 0$ is a user-defined regularization hyperparameter. We assume that the negative log-likelihood is strictly convex. This assumption is satisfied for instance in the Gaussian linear model or logistic regression. In this case, it is necessary and sufficient that the solutions $\hat{\vartheta}^\lambda$ to (6.12) satisfy the following

Karush–Kuhn–Tucker (KKT) conditions

$$\begin{cases} \mathbf{X}^\top \left(Y - \sigma(\mathbf{X}\hat{\vartheta}^\lambda)\right) = \lambda \widehat{S} & \text{(6.13a)} \\ \widehat{S}_k = \operatorname{sign}(\hat{\vartheta}^\lambda_k) \qquad \text{if } \hat{\vartheta}^\lambda_k \neq 0 & \text{(6.13b)} \\ \widehat{S}_k \in [-1, 1] \qquad \text{if } \hat{\vartheta}^\lambda_k = 0 & \text{(6.13c)} \end{cases}$$

Proposition 6.1 shows that there exists only one vector of signs $\widehat{S} \in \mathbb{R}^d$ such that $(\hat{\vartheta}^\lambda, \hat{S})$ satisfies the KKT conditions for some $\hat{\vartheta}^\lambda \in \mathbb{R}^d$. The proof of Proposition 6.1 can be found in Section 6.8.1.

Proposition 6.1. *Let $Y \in \mathcal{Y}^N$ and let the partition function ξ be strictly convex. Then, there exists a unique $\widehat{S}(Y)$ such that for any couple $(\hat{\vartheta}^\lambda, \hat{S})$ satisfying the KKT conditions (cf. Eq.(6.13)) it holds that $\hat{S} = \widehat{S}(Y)$. Furthermore, one has*

$$\widehat{S}(Y) := \frac{1}{\lambda} \mathbf{X}^\top (Y - \sigma(\mathbf{X}\hat{\vartheta}^\lambda)),$$

where $\hat{\vartheta}^\lambda$ is any solution of the generalized linear Lasso as defined in (6.12).

We define the *equicorrelation set* as

$$\widehat{M}(Y) := \{k \in [d] \mid |\widehat{S}_k(Y)| = 1\}.$$

In the following, we will identify the equicorrelation set and the set of predictors with nonzero coefficients $\{k \in [d] \mid \hat{\vartheta}^\lambda_k \neq 0\}$, also called "selected" model. Since $|\widehat{S}_k(Y)| = 1$ for any $\hat{\vartheta}^\lambda_k \neq 0$, the equicorrelation set does in fact contain all predictors with nonzero coefficients, although it may also include some predictors with zero coefficients. However, we work in this chapter with Assumption 6, ensuring that the equicorrelation set is precisely the set of predictors with nonzero coefficients.

Assumption 6. *Problem (6.12) is non degenerate: $\widehat{S}(Y) \in \operatorname{relint} \partial \|\cdot\|_1$, where* relint *denotes the relative interior.*

Let us highlight that this assumption has already been used in the context of GLMs [cf. Massias et al., 2020, Assumption 8], and is common in works on support identification (cf. Candes and Recht [2013], Vaiter et al. [2015]).
For any set of indexes $M \subseteq [d]$ with cardinality s, the set of target parameters induced by the support M is $\mathbb{R}^s$. We aim at making inference conditionally on the *selection event* E_M defined as

$$E_M := \left\{ Y \in \mathcal{Y}^N \mid \widehat{M}(Y) = M \right\}, \tag{6.14}$$

namely, the set of all observations Y that induced the same equicorrelation set M with the generalized linear lasso.

6.2.2 Characterization of the selection event

Following the approach of Lee et al. [2016], given some $M \subseteq [d]$ with $|M| = s$ and $S_M \in \{-1, +1\}^s$, we first characterize the event

$$E_M^{S_M} := \{Y \in E_M \mid \widehat{S}_M(Y) = S_M\}, \tag{6.15}$$

and we obtain E_M as a corollary by taking a union over all possible vectors of signs S_M. Proposition 6.2 gives a first description of $E_M^{S_M}$ and its proof is postponed to Section 6.8.2.

Proposition 6.2. *Let us consider $M \subseteq [d]$ with $|M| = s$ and $S_M \in \{-1, +1\}^s$. It holds*

$$E_M^{S_M} = \Big\{ Y \in \mathcal{Y}^N \mid \exists \theta \in \mathbb{R}^s \text{ s.t. } (i)\ \mathbf{X}_M^\top (Y - \sigma(\mathbf{X}_M \theta)) = \lambda S_M \tag{6.16}$$
$$(ii)\ \operatorname{sign}(\theta) = S_M$$
$$(iii)\ \left\| \mathbf{X}_{-M}^\top (Y - \sigma(\mathbf{X}_M \theta)) \right\|_\infty < \lambda \Big\},$$

were $\mathbf{X}_M \in \mathbb{R}^{N \times s}$ (resp. $\mathbf{X}_{-M} \in \mathbb{R}^{N \times (d-s)}$) is the submatrix obtained from $\mathbf{X}$ by keeping the columns indexed by M (resp. its complement).

With Proposition 6.1, we proved the uniqueness of the vector of signs satisfying the KKT conditions as soon as ξ is strictly convex. By considering additionally that $\mathbf{X}_M$ has full column rank, we claim that there exists a unique $\theta \in \mathbb{R}^s$ that satisfies the condition (i) in the definition of the selection event $E_M^{S_M}$ (see Eq.(6.16)). This statement will be a direct consequence of Proposition 6.3 (proved in Section 6.8.3) which ensures that the map Ξ arising in Eq.(6.16) and defined by

$$\begin{aligned} \Xi : \mathbb{R}^s &\to \mathbb{R}^s \\ \theta &\mapsto \mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta) \end{aligned} \tag{6.17}$$

is a $\mathcal{C}^m$-diffeomorphism whose inverse is denoted by Ψ.

Proposition 6.3. *We consider that the partition function ξ is strictly convex and we further assume that the set $M \subseteq [d]$ is such that $\mathbf{X}_M$ has full column rank. Then Ξ is a $\mathcal{C}^m$-diffeomorphism from $\mathbb{R}^s$ to $\mathrm{Im}(\Xi) = \{\mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta) \mid \theta \in \mathbb{R}^s\}$.*

Using Propositions 6.2 and 6.3, we are able to provide a new description of the selection event $E_M^{S_M}$ which can be understood as the counterpart of [Lee et al., 2016, Proposition 4.2].

Theorem 6.4. *Suppose that ξ is strictly convex. Given some $M \subseteq [d]$ with cardinal s such that $\mathbf{X}_M$ has full column rank and $S_M \in \{-1,1\}^s$, it holds*

$$\begin{aligned} E_M^{S_M} = \Big\{ Y \in \mathcal{Y}^N \mid \text{s.t. } \rho = -\lambda S_M + \mathbf{X}_M^\top Y \text{ satisfies} \\ (a)\ \rho \in \mathrm{Im}(\Xi) \\ (b)\ \mathrm{Diag}(S_M)\Psi(\rho) \geq 0 \\ (c)\ \left\| \mathbf{X}_{-M}^\top \left(Y - \sigma(\mathbf{X}_M \Psi(\rho)) \right) \right\|_\infty < \lambda \Big\}. \end{aligned} \tag{6.18}$$

Remark. In the linear model, $\Xi : \theta \mapsto \mathbf{X}_M^\top \mathbf{X}_M \theta$ has full rank and thus condition (a) from Eq.(6.18) always holds.

6.2.3 Which parameters can be inferred?

Once a model has been selected, two different modeling assumptions are generally considered when we derive post-selection inference procedures, see for instance [Fithian et al., 2014, Section 4]. This choice appears to be essential since it determines the parameters on which inference is conducted. In the following, we consider the mean value

$$\pi^* := \mathbb{E}_{\vartheta^*}[Y] = \sigma(\mathbf{X}\vartheta^*)\,. \tag{6.19}$$

Note that π^* allows to define the Bayes predictor in the logistic or the linear model. As presented in Fithian et al. [2014], the analyst should decide to work either under the so-called *saturated model* or the *selected model*. In the following, we discuss these concepts for arbitrary GLMs and Table 6.3 summarizes the key concepts.

The (weak) selected model: Parameter inference. In the *weak selected model*, we consider that the data have been sampled from the GLM (cf. Eq.(6.11)) and we assume that the selected model M is such that

$$\mathbf{X}_M^\top \sigma(\mathbf{X}\vartheta^*) \in \mathrm{Im}(\Xi)\,, \tag{6.20}$$

and recall that $\mathbf{X}_M^\top \pi^* = \mathbf{X}_M^\top \mathbb{E}_{\vartheta^*}[Y] = \mathbf{X}_M^\top \sigma(\mathbf{X}\vartheta^*)$. This is equivalent to state that there exists some vector $\theta^* \in \mathbb{R}^s$ satisfying

$$\mathbf{X}_M^\top \pi^* = \Xi(\theta^*)\,,$$

and recall that $\Xi(\theta^*) = \mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta^*)$. In this framework, we have the possibility to make inference on the parameter vector $\theta^* := \Psi(\mathbf{X}_M^\top \pi^*)$ itself.
In the *selected model*, we replace the condition from Eq.(6.20) by the stronger assumption that there exists $\theta^* \in \mathbb{R}^s$ such that

$$\mathbf{X}_M \theta^* = \mathbf{X}\vartheta^*. \tag{6.21}$$

This assumption is always satisfied for the global null hypothesis $\vartheta^* = 0$ for which the aforementioned condition holds with $\theta^* = 0$.

Model	Selected	Weak selected	Saturated
Assumption	$\sigma^{-1}(\pi^*) \in \mathrm{Im}(\mathbf{X}_M)$	$\mathbf{X}_M^\top \pi^* \in \mathrm{Im}(\Xi)$	None
Statistic of interest	$\Psi(\mathbf{X}_M^\top Y)$	$\Psi(\mathbf{X}_M^\top Y)$	$\mathbf{X}_M^\top Y$
Inferred parameter	$\theta^* \in \mathbb{R}^s$ s.t. $\pi^* = \sigma(\mathbf{X}_M \theta^*)$	$\theta^* \in \mathbb{R}^s$ s.t. π^* and $\sigma(\mathbf{X}_M\theta^*)$ have the same projections on the column span of $\mathbf{X}_M$	$\mathbf{X}_M^\top \pi^*$

Table 6.3: Once a model has been selected, we may infer some parameters assuming one of the three modeling: selected model, weak selected model, and saturated model respectively based on the assumptions described in the first row. In this case, inference on the quantities described on the third row can be done from the statistic described in the second row.

The saturated model: Mean value inference. The assumption from Eq.(6.20) or (6.21) can be understood as too restrictive since the analyst can never check in practice that this condition holds, except for the global null. This is the reason why one may prefer to consider the so-called *saturated model* where we only assume that the data have been sampled from the GLM.
In this case it remains meaningful to provide post-selection inference procedure for transformation of π^*. A typical choice is to consider linear transformation of π^* and among them, one may focus specifically on transformation of $\mathbf{X}_M^\top \pi^*$. This choice is motivated by remarking that this quantity characterizes the first order optimality condition for the unpenalized MLE $\widehat{\theta}$ for the design matrix $\mathbf{X}_M$ through $\mathbf{X}_M^\top Y = \Xi(\widehat{\theta})$, or by considering the example of linear model (as presented below).

The example of the linear model. Note that in linear regression, $\sigma = \mathrm{Id}$ and $\Psi : \rho \mapsto \left(\mathbf{X}_M^\top \mathbf{X}_M\right)^{-1}\rho$. Hence, Eq.(6.20) is equivalent to Eq.(6.21) meaning that the selected and the weak selected models coincide. Moreover, in both the saturated and the selected models, we aim at making inference on transformations of $\Psi(\mathbf{X}_M^\top \pi^*) = \mathbf{X}_M^+ \pi^*$ (where $\mathbf{X}_M^+$ is the pseudo-inverse of $\mathbf{X}_M$). While in the (weak) selected model, this quantity corresponds to the parameter vector θ^* satisfying $\pi^* = \mathbf{X}_M\theta^*$, in the saturated model, it corresponds to the best linear predictor in the population for design matrix $\mathbf{X}_M$ in the sense of the squared L^2-norm.

6.2.4 Inference procedures

We provide a general approach to obtain asymptotically valid PSI methods, both in the saturated and the selected models. The proposed PSI methods rely on two key ingredients, namely conditional Central Limit Theorems (CLTs) and conditional sampling. Before describing our selective inference procedures, let us set the rigorous framework for which we provide our conditional CLTs.

Preliminaries. Given a non-decreasing sequence of positive integers $(d_N)_{N\in\mathbb{N}}$ converging to $d_\infty \in \mathbb{N}\cup\{+\infty\}$, we consider for any N a matrix $\mathbf{X}^{(N)} \in \mathbb{R}^{N\times d_N}$ and a vector $[\vartheta^*]^{(N)} \in \mathbb{R}^{d_N}$. Let $s \in [d_1, d_\infty]\cap\mathbb{N}$ be a fixed and finite integer and let us consider for any N a set $M^{(N)} \subseteq [d_N]$ with cardinality s. Considering further a sequence of positive reals $(\lambda^{(N)})_{N\in\mathbb{N}}$, we define $E_M^{(N)}$ as the selection event corresponding to the tuple $(\lambda^{(N)}, \mathbf{X}^{(N)}, M^{(N)})$ meaning that

$$E_M^{(N)} := \bigcup_{S_M\in\{\pm1\}^N} \left[E_M^{S_M}\right]^{(N)},$$

where $\left[E_M^{S_M}\right]^{(N)}$ is the set given by Eq.(6.18) when one uses the regularization parameter $\lambda^{(N)}$, the design matrix $\mathbf{X}^{(N)}$ and the set of active variables $M^{(N)}$. Considering that the $\mathcal{Y}^N$-valued random vector Y is distributed according to

$$\overline{\mathbb{P}}_{\pi^*}^{(N)}(Y) \propto \mathbb{1}_{Y\in E_M^{(N)}} \mathbb{P}_{\pi^*}^{(N)}(Y), \tag{6.22}$$

with $\mathbb{P}_{\pi^*}^{(N)} := \mathbb{P}_{\sigma(\mathbf{X}^{(N)}[\vartheta^*]^{(N)})}$, the cornerstone of our methods consists in proving a CLT for $\big[\mathbf{X}_{M^{(N)}}^{(N)}\big]^\top Y$ in the saturated model, see Eq.(6.23) (resp. a CLT for the conditional MLE $\Psi(\mathbf{X}_M^\top Y)$ in the selected model, see Eq.(6.24)).

In Section 6.5, we consider the specific case of the logistic regression and we establish conditional CLTs of the form given by Eqs.(6.23) and (6.24). The proofs of our CLTs rely on triangular arrays of dependent random vectors of the form $(\xi_{i,N})_{i\in[N]}$. For any fixed N and any $i \in [N]$, $\xi_{i,N}$ is a random vector in $\mathbb{R}^s$ which can be written as a function of the deterministic quantities $\lambda^{(N)}$, $\mathbf{X}^{(N)}$, $M^{(N)}$ and of the random variable Y with probability distribution $\overline{\mathbb{P}}_{\pi^*}^{(N)}$. In Section 6.5, we provide conditions ensuring that the rows of the triangular system $\big((\xi_{i,N})_{i\in[N]},\ N \in \mathbb{N}\big)$ satisfy some Lindeberg's condition.

With this detailed framework now established, we will take the liberty in the remainder of this chapter of adopting certain abuses of notation for the sake of readability. The notational clutter will be in particular reduced by forgetting to specify the dependence on N meaning that we will simply refer to $\mathbf{X}^{(N)}$, $M^{(N)}$, d_N, $[\vartheta^*]^{(N)}$, $\overline{\mathbb{P}}_{\pi^*}^{(N)}, \dots$ as $\mathbf{X}$, M, d, ϑ^*, $\overline{\mathbb{P}}_{\pi^*}, \dots$. Nevertheless, let us stress again that the integer s is fixed and does not depend on N in our work.

Conditional CLTs. The cornerstone of our method is to establish conditional CLTs. More precisely, considering that Y is distributed according to $\overline{\mathbb{P}}_{\pi^*} = \mathbb{P}_{\pi^*}(\cdot \,|\, E_M)$, we aim at providing conditions ensuring that

- in the saturated model,

$$\overline{G}_N(\pi^*)^{-1/2}(\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi^*}) \xrightarrow[N\to\infty]{(d)} \mathcal{N}(0, \mathrm{Id}_s), \tag{6.23}$$

  for some $\overline{G}_N(\pi^*) \in \mathbb{R}^{s\times s}$ and $\overline{\pi}^{\pi^*} \in \mathbb{R}^N$ depending only on π^* and E_M,

- in the selected model where $\mathbf{X}\vartheta^* = \mathbf{X}_M\theta^*$,

$$\overline{V}_N(\theta^*)^{1/2}(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta^*)) \xrightarrow[N\to\infty]{(d)} \mathcal{N}(0, \mathrm{Id}_s), \tag{6.24}$$

  for some $\overline{V}_N(\theta^*) \in \mathbb{R}^{s\times s}$ and $\overline{\theta}(\theta^*) \in \mathbb{R}^s$ depending only on θ^* and E_M.

In the case of logistic regression, we give in Section 6.5 conditions ensuring that the CLTs from Eqs.(6.23) and (6.24) hold.

Idyllic selective inference. Using the above mentioned conditional CLTs, one can obtain confidence regions (CRs) with asymptotic level $1-\alpha$ (for some $\alpha \in (0,1)$) conditional on E_M as follows,

- in the saturated model, the CR for π^* is defined as

$$\{\pi \,|\, \|\overline{G}_N(\pi)^{-1/2}(\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi})\|_2^2 \le \chi^2_{s,1-\alpha}\}, \tag{6.25}$$

- in the selected model, the CR for θ^* is defined as

$$\{\theta \,|\, \|\overline{V}_N(\theta)^{1/2}(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta))\|_2^2 \le \chi^2_{s,1-\alpha}\}, \tag{6.26}$$

where $\chi^2_{s,1-\alpha}$ is the quantile of order $1-\alpha$ of the χ^2 distribution with s degrees of freedom. Obviously, covering the whole space to obtain in practice the CRs from Eqs.(6.25) and (6.26) is out of reach and one could use a discretization of a bounded domain to bypass this limitation. A more involved issue is that $\overline{G}_N(\pi)$ and $\overline{\pi}^{\pi}$, (resp. $\overline{V}_N(\theta)$ and $\overline{\theta}(\theta)$) can be written as expectations with respect to the conditional distribution $\overline{\mathbb{P}}_\pi$ (resp. $\overline{\mathbb{P}}_\theta$). If closed-form expressions most of time do not exist, one can estimate the latter quantities by sampling from $\overline{\mathbb{P}}_\pi$ (resp. $\overline{\mathbb{P}}_\theta$). Depending on the studied GLM, this task may be expensive and the proposed grid-based approaches to get CRs would be unusable in practice due to the curse of dimensionality. In the next subparagraph, we propose an alternative method to overcome this issue.

Confidence regions in practice. Table 6.4 gives the main ideas allowing us to obtain a confidence region for π^* (resp. θ^*) in the saturated (resp. selected) model. While the blue terms are small with high probability for N large enough thanks to the previous established conditional CLTs, the red terms motivate us to choose our estimate $\pi^\star$ (resp. $\theta^\star$) among the minimizers of the map $\pi \mapsto \|\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^\pi\|_2$ (resp. $\theta \mapsto \|\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta)\|_2$). This way, we circumvent the curse of dimensionality of the above mentioned idyllic approach to obtain CRs. Nevertheless, the given CRs involve quantities that are unknown in practice such as the constant κ_1 (resp. κ_2) in Table 6.4 that encodes the (local) Lipschitz continuity of the inverse of the map $\pi \mapsto \overline{\pi}^\pi$ (resp. $\theta \mapsto \overline{\theta}(\theta)$). Note that the term $\|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star})\|_2$ arising in the CR for the saturated model illustrates that we only control what occurs on the span of the columns of $\mathbf{X}_M$.

Saturated model	$\forall \pi^\star,\ \|\pi^* - \pi^\star\|_2 \lesssim \kappa_1 \|\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star}\|_2$ $\leq \kappa_1 \left\{ \|\mathrm{Proj}_{\mathbf{X}_M}(\overline{\pi}^{\pi^*} - Y)\|_2 + \|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^\star})\|_2 + \|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star})\|_2 \right\}$
Selected model	$\forall \theta^\star,\ \|\theta^* - \theta^\star\|_2 \lesssim \kappa_2 \|\overline{\theta}(\theta^*) - \overline{\theta}(\theta^\star)\|_2$ $\leq \kappa_2 \left\{ \|\overline{\theta}(\theta^*) - \Psi(\mathbf{X}_M^\top Y)\|_2 + \|\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta^\star)\|_2 \right\}$

Table 6.4: Confidence intervals.

In the case of logistic regression, we present in Section 6.6 with full details our selective inference procedures with theoretical guarantees.

Conditional sampling: A Monte-Carlo approach for hypothesis-testing. We consider hypothesis tests with pointwise nulls as presented in Table 6.5. One can then compute estimate $\widetilde{G}_N(\pi_0^*), \widetilde{\pi}^{\pi_0^*}$ (resp. $\widetilde{V}_N(\theta_0^*), \widetilde{\theta}(\theta_0^*)$) of the unknown quantities $\overline{G}_N(\pi_0^*)$, $\overline{\pi}^{\pi_0^*}$ (resp. $\overline{V}_N(\theta_0^*)$, $\overline{\theta}(\theta_0^*)$) by sampling from the conditional null distribution $\overline{\mathbb{P}}_{\pi_0^*}$ (resp. $\overline{\mathbb{P}}_{\theta_0^*}$ in the selected model). Using a Monte-Carlo approach with the CLTs from Eqs.(6.23) and (6.24), one can derive testing procedures that are asymptotically correctly calibrated.

	Null and alternative	Distributed *approximately* as $\mathcal{N}(0, \mathrm{Id}_s)$ under $\mathbb{H}_0$
Saturated model	$\mathbb{H}_0 : \{\pi^* = \pi_0^*\}$, $\mathbb{H}_1 : \{\pi^* \neq \pi_0^*\}$	$\widetilde{G}_N(\pi_0^*)^{-1/2}(\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \widetilde{\pi}^{\pi_0^*})$
Selected model	$\mathbb{H}_0 : \{\theta^* = \theta_0^*\}$, $\mathbb{H}_1 : \{\theta^* \neq \theta_0^*\}$	$\widetilde{V}_N(\theta_0^*)^{1/2}(\Psi(\mathbf{X}_M^\top Y) - \widetilde{\theta}(\theta_0^*))$

Table 6.5: Hypothesis testing.

In the case of logistic regression, we rely on a gradient alignment viewpoint of the selection event to provide in Section 6.4 an algorithm which allows us to sample from $\overline{\mathbb{P}}_{\pi^*}$ given any π^*. In Section 6.6, we present our hypothesis tests in both the saturated and the selected models with theoretical guarantees.

6.2.5 Related works

In the Gaussian linear model with a known variance, the distribution of the linear transformation $\eta^\top Y$ (with $\eta^\top = e_k^\top \mathbf{X}_M^+$) is a truncated Gaussian conditionally on $E_M^{S_M}$ and $\mathrm{Proj}_\eta^\perp(Y)$. This explicit formulation of the conditional distribution allows to conduct exact post-selection inference procedures [cf. Fithian et al., 2014, Section 4]. However, when the noise is assumed to be Gaussian with an unknown variance, one needs to also condition on $\|Y\|^2$ which leaves insufficient information about θ_k^* to carry out a meaningful test in the saturated model [cf. Fithian et al., 2014, Section 4.2].
Outside of the Gaussian linear model, there is little hope to obtain a useful exact characterization of the conditional distribution of some transformation of $\mathbf{X}_M^\top Y$. In the following, we sketch a brief review of this literature, see references therein for further works on this subject.

- Linear model but non-Gaussian errors.
  Let us mention for example Tian and Taylor [2017], Tibshirani et al. [2018] where the authors

consider the linear model but relaxed the Gaussian distribution assumption for the error terms. They prove that the response variable is asymptotically Gaussian so that applying the well-oiled machinery from Lee et al. [2016] gives asymptotically valid post-selection inference methods.

- GLM with Gaussian errors.
  Shi et al. [2020] consider generalized linear models with Gaussian noise and can then immediately apply the polyhedral lemma to the appropriate transformation of the response.

We classify existing works with Table 6.6.

Noise	Linear Model	GLM
Gaussian	Lee et al. [2016]	Shi et al. [2020]
Non-Gaussian	Tian and Taylor [2017] and Tibshirani et al. [2018]	Our work and Taylor and Tibshirani [2018]

Table 6.6: Positioning of our contributions among some pioneering works on PSI in GLMs.

One important challenge that remains so far only partially answered is the case of GLMs without Gaussian noise, such as in logistic regression. In Fithian et al. [2014], the authors derive powerful unbiased selective tests and confidence intervals among all selective level-α tests for inference in exponential family models after arbitrary selection procedures. Nevertheless, their approach is not well-suited to account for discrete response variable as it is the case in logistic regression. In Section 6.3 of the former paper, the authors rather encourage the reader to make use of the procedure proposed by Taylor and Tibshirani [2018] in such context. Both our work and Taylor and Tibshirani [2018] are tackling the problem of post selection inference in the logisitic model. Nevertheless, the proposed methods rely on different paradigms and we explain in Section 6.3 this difference in perspectives.

6.2.6 Contributions and organization of this chapter

Working with an arbitrary GLM (Sec.6.2 and 6.3).

1. We provide a new formulation of the selection event in GLMs shedding light on the $\mathcal{C}^m$-diffeomorphism Ψ that carries the geometric information of the problem (cf. Theorem 6.4). Ψ allows us to define rigorously the notions of selected/saturated models for arbitrary GLM (cf. Sec.6.2.3).
2. We provide a new perspective on post-selection inference in the selected model for GLMs through the conditional MLE approach of which Ψ is a key ingredient (cf. Sec.6.3).
3. We introduce the C-cube conditions that are sufficient conditions in GLMs to obtain valid post-selection inference procedures in the selected model based on the conditional MLE approach (cf. Sec.6.3).

Considering the Sparse Logistic Regression (SLR) (from Sec.6.4).

4. Under some assumptions, we prove that the C-cube conditions hold for the SLR and we conduct simulations to support our results.
5. We also derive asymptotically valid PSI methods in the saturated model for the SLR.
6. We provide an extensive comparison between our work and the heuristic from Taylor and Tibshirani [2018] which is currently considered the best to use in the context of SLR [cf. Fithian et al., 2014, Section 6.3], as far as we know.

Outline. In Section 6.3, we introduce the conditional MLE approach to tackle PSI in the selected model and we stress the difference with the debiasing method from Taylor and Tibshirani [2018]. From Section 6.4, we focus specifically on the SLR. In Section 6.4, we rely on a *gradient-alignment* viewpoint on the selection event to design a simulated annealing algorithm which is proved–for an appropriate cooling scheme–to provide iterates whose distribution is asymptotically uniform on the selection event. In

Section 6.5, we provide two conditional central limit theorems that would be key theoretical ingredients for our PSI methods presented in Section 6.6. In Section 6.6.1, we give PSI procedures in the selected model while in Section 6.6.2 we focus on the saturated model. In Section 6.7, we present the results of our simulations.

Notations. For any set of indexes $M \subseteq [d] := \{1, \dots, d\}$ and any vector v, we denote by v_M the subvector of v keeping only the coefficients indexed by M, namely $v_M = (v_k)_{k \in M}$. Analogously, v_{-M} will refer to the subvector $(v_k)_{k \notin M}$. $|M|$ denotes the cardinality of the finite set M. For any $x \in \mathbb{R}^d$, $\|x\|_\infty := \sup_{i \in [d]} |x_i|$ and for any $p \in [1, \infty)$, $\|x\|_p^p := \sum_{i \in [d]} x_i^p$. For any $A \in \mathbb{R}^{d \times p}$, we define the Frobenius norm of A as $\|A\|_F := (\sum_{i \in [d], j \in [p]} A_{i,j}^2)^{1/2}$ and the operator norm of A as $\|A\| := \sup_{x \in \mathbb{R}^p, \|x\|_2 = 1} \|Ax\|_2$. We further denote by A^+ the pseudo-inverse of A. Considering that A is a symmetric matrix, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ will refer respectively to the minimal and the maximal eigenvalue of A. $\odot$ denotes the Hadamard product namely for any $A, B \in \mathbb{R}^{d \times p}$, $A \odot B := (A_{i,j} B_{i,j})_{i \in [d], j \in [p]}$. By convention, when a function with real valued arguments is applied to a vector, one need to apply the function entrywise. $\mathrm{Id}_d \in \mathbb{R}^{d \times d}$ will refer to the identity matrix and $\mathcal{N}(\mu, \Sigma)$ will denote the multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ. For any $x \in \mathbb{R}^d$, $R > 0$ and for $p \in [1, \infty]$, we define $\mathbb{B}_p(x, R) = \{z \in \mathbb{R}^d \,|\, \|z\|_p \leq R\}$.
Let us finally recall that given some set of selected variables $M \subseteq [d]$ with $s := |M|$ and some $\vartheta^* \in \mathbb{R}^d$, we denote by $\overline{\mathbb{P}}_{\pi^*}$ the distribution of Y conditional on E_M, namely

$$\overline{\mathbb{P}}_{\pi^*}(Y) \propto \mathbb{1}_{Y \in E_M} \mathbb{P}_{\pi^*}(Y),$$

$\pi^* = \sigma(\mathbf{X}\vartheta^*)$ and where $\propto$ means equal up to a normalization constant (cf. Eq.(6.22)). By assuming that ξ is strictly convex, one can compute $\mathbf{X}\vartheta^*$ from π^*, allowing us to denote equivalently $\mathbb{P}_{\pi^*} \equiv \mathbb{P}_{\vartheta^*}$ with an abuse of notation. In the selected model with $\theta^* \in \Theta_M$ satisfying Eq.(6.21), we will also denote $\mathbb{P}_{\pi^*} \equiv \mathbb{P}_{\theta^*}$.

6.3 Regularization bias and conditional MLE

In this section, we wish to emphasize the different nature of our approach and that of Taylor and Tibshirani [2018] which we consider as the more relevant point of comparison, to the best of our knowledge. While we rely on a conditional MLE viewpoint, the former paper consider a debiasing approach.

- *The debiasing approach*
  ℓ_1-penalization induced a soft-thresholding bias and one can first try to modify the solution of the penalized GLM $\hat{\vartheta}^\lambda$ to approximate the unconditional MLE of the GLM using only the features in the selected support M by some vector $\underline{\theta}$. Provided that we work with a *correctly specified model* M*–i.e.*, one that contains the true support $\{j \in [d] \,|\, \vartheta_j^* \neq 0\}$–standard results ensure that the unconditional MLE is asymptotically normal, asymptotically efficient and centered at ϑ_M^*. If one can show that the selection event only involve polyhedral constraints on a linear transformation $\eta^\top \underline{\theta}$ of the debiased vector $\underline{\theta}$, the conditional distribution of $\eta^\top \underline{\theta}$ would be a truncated Gaussian. This is the approach from Taylor and Tibshirani [2018] that we detail in Section 6.3.1.

- *The conditional MLE viewpoint*
  In this chapter, we follow a different route: one can grasp the nettle by studying directly the properties of the unpenalized conditional MLE.

6.3.1 Selective inference through debiasing

The idea behind the method proposed by Taylor and Tibshirani [2018] is that we need two key elements to mimic the approach from Lee et al. [2016] proposed in the linear model with Gaussian errors:

- A statistic $T(Y)$ converging in distribution to a Gaussian distribution with a mean involving the parameter of interest;

- A selection event that can be written as a union of polyhedra with respect to $\eta^\top T(Y)$ for some vector η.

In practice, a solution of the generalized linear Lasso (cf. Eq.(6.12)) can be approximated using the Iteratively Reweighted Least Squares (IRLS). Defining

$$W(\vartheta) = \nabla^2_\eta \mathcal{L}_N(\eta)\big|_{\eta=\mathbf{X}\vartheta} = \mathrm{Diag}(\sigma'(\mathbf{X}\vartheta)),$$
$$\text{and } z(\vartheta) = \mathbf{X}\vartheta - [W(\vartheta)]^{-1}\nabla_\eta \mathcal{L}_N(\eta)\big|_{\eta=\mathbf{X}\vartheta} = \mathbf{X}\vartheta + [W(\vartheta)]^{-1}(Y - \sigma(\mathbf{X}\vartheta)),$$

the IRLS algorithm works as follows.

1: Initialize $\vartheta_c = 0$.
2: Compute $W(\vartheta_c)$ and $z(\vartheta_c)$.
3: Update the current value of the parameters with
$$\vartheta_c \leftarrow \arg\min_\vartheta \frac{1}{2}(z(\vartheta_c) - \mathbf{X}\vartheta)^\top W(\vartheta_c)(z(\vartheta_c) - \mathbf{X}\vartheta) + \lambda\|\vartheta\|_1.$$
4: Repeat steps 2. and 3. until convergence.

If the IRLS has converged, we end up with a solution $\hat{\vartheta}^\lambda$ of Eq.(6.12) and, for $M = \{j \in [d] \,|\, \hat{\vartheta}^\lambda_j \neq 0\}$, the active block of stationary conditions (Eq. (6.16) (i)) can be written as

$$\mathbf{X}_M^\top W \left\{ z - \mathbf{X}_M \hat{\vartheta}^\lambda_M \right\} = \lambda S_M,$$

where $W = W(\hat{\vartheta}^\lambda)$, $z = z(\hat{\vartheta}^\lambda)$ and $S_M = \mathrm{sign}(\hat{\theta}^\lambda_M)$. The solution $\hat{\vartheta}^\lambda_M$ should be understood as a biased version of the unpenalized MLE $\widehat{\theta}$ obtained by working on the support M, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^s} \sum_{i=1}^N \xi(\langle \mathbf{X}_{i,M}, \theta \rangle) - \langle y_i \mathbf{X}_{i,M}, \theta \rangle.$$

If we work with a *correctly specified model* M*–i.e.,* one that contains the true support $\{j \in [d] \,|\, \vartheta^*_j \neq 0\}$– then it follows from standard results that the MLE $\widehat{\theta}$ is a consistent and asymptotically efficient estimator of ϑ^*_M (see e.g. [Van der Vaart, 2000, Theorem 5.39]). A natural idea consists in debiasing the vector of parameters ϑ^λ_M in order to get back to the parameter $\widehat{\theta}$ and to use its nice asymptotic properties for inference. We thus consider

$$\underline{\theta} = \vartheta^\lambda_M + \lambda \left(\mathbf{X}_M^\top W \mathbf{X}_M\right)^{-1} S_M,$$

so that $\underline{\theta}$ satisfies

$$\mathbf{X}_M^\top W \{z - \mathbf{X}_M \underline{\theta}\} = 0. \tag{6.27}$$

If one replaces W and z in Eq.(6.27) by $W(\underline{\vartheta})$ and $z(\underline{\vartheta})$ (with the obvious notation that $\underline{\vartheta}_M = \underline{\theta}$ and $\underline{\vartheta}_{-M} = 0$), Eq.(6.27) corresponds to the stationarity condition of the unpenalized MLE for the generalized linear regression using only the features in M.
Hence, Taylor and Tibshirani [2018] propose to treat the debiased parameters $\underline{\theta}$ has asymptotically normal centered at ϑ^*_M with covariance matrix $\left(\mathbf{X}_M^\top W(\vartheta^*) \mathbf{X}_M\right)^{-1}$. Since ϑ^* is unknown, they use a Monte-Carlo estimate and replace $W(\vartheta^*)$ by $W(\hat{\vartheta}^\lambda)$ in the Fisher information matrix. By considering that $\vartheta^* = N^{-1/2}\beta^*$ where each entry of β^* is independent of N, they claim that the selection event $E_M^{S_M}$ can be asymptotically approximated by

$$\mathrm{Diag}(S_M) \left(\underline{\theta} - \lambda \left(\mathbf{X}_M^\top W \mathbf{X}_M\right)^{-1} S_M \right) \geq 0.$$

Hence, to derive post-selection inference procedure, they apply the polyhedral lemma to the limiting distribution of $N^{1/2}\underline{\theta}$, with M and S_M fixed.

6.3.2 Selective inference through conditional MLE

We change of paradigm and we directly work with the conditional distribution. The conditional distribution given $Y \in E_M$ is a conditional exponential family with the same natural parameters and

sufficient statistics but different support and normalizing constant:

$$\overline{\mathbb{P}}_\theta(Y) \propto \mathbb{1}_{E_M}(Y) \prod_{i=1}^N h_\theta(y_i) \exp\left\{ \frac{y_i \mathbf{X}_{i,M}\theta - \xi(\mathbf{X}_{i,M}\theta)}{v} \right\},$$

where the symbol $\propto$ means "proportional to". When $E_M = \mathcal{Y}^N$ (*i.e.,* when there is no conditioning), we will simply denote $\overline{\mathbb{P}}_\theta$ by $\mathbb{P}_\theta$. In the following we will denote by $\overline{\mathbb{E}}_\theta$ (resp. $\mathbb{E}_\theta$) the expectation with respect to $\overline{\mathbb{P}}_\theta$ (resp. $\mathbb{P}_\theta$). We want to conduct inference on θ^* (from Eq.(6.21)) based on the conditional and unpenalized MLE computed on the selected model M, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^s} \mathcal{L}_N(\theta, Z^M), \tag{6.28}$$

where $Z^M = (Y, \mathbf{X}_M)$ and where Y is distributed according to $\overline{\mathbb{P}}_{\theta^*}$. We aim at proving a Central Limit Theorem for $\widehat{\theta}$. A natural candidate for the mean of the asymptotic Gaussian distribution of $\widehat{\theta}$ is the minimizer of the conditional expected negative log-likelihood defined by

$$\overline{\theta}(\theta^*) \in \arg\min_{\theta \in \mathbb{R}^s} \overline{\mathbb{E}}_{\theta^*}\left[\mathcal{L}_N(\theta, Z^M)\right], \tag{6.29}$$

which is the minimizer of the conditional risk $\theta \mapsto \overline{\mathbb{E}}_{\theta^*}\left[\mathcal{L}_N(\theta, Z^M)\right]$.
In the following, when there is no ambiguity we will simply denote $\overline{\theta}(\theta^*)$ by $\overline{\theta}$. Hence, denoting

$$L_N(\theta, Z^M) = \frac{\partial \mathcal{L}_N}{\partial \theta}(\theta, Z^M) \quad \text{and} \quad \overline{L}_N(\theta, \mathbf{X}_M) = \overline{\mathbb{E}}_{\theta^*}\left[\frac{\partial \mathcal{L}_N}{\partial \theta}(\theta, Z^M)\right],$$

it holds that the conditional unpenalized MLE $\widehat{\theta}$ and the minimizer $\overline{\theta}$ of the conditional risk satisfy the first order condition

$$\begin{aligned} L_N(\widehat{\theta}, Z^M) = 0 \quad i.e. \quad \mathbf{X}_M^\top (Y - \pi^{\widehat{\theta}}) = 0 \Leftrightarrow \widehat{\theta} = \Psi(\mathbf{X}_M^\top Y), \\ \text{and } \overline{L}_N(\overline{\theta}, \mathbf{X}_M) = 0 \quad i.e. \quad \mathbf{X}_M^\top (\overline{\pi}^{\theta^*} - \pi^{\overline{\theta}}) = 0 \Leftrightarrow \overline{\theta} = \Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta^*}), \end{aligned} \tag{6.30}$$

where $\pi^\theta = \mathbb{E}_\theta[Y] = \sigma(\mathbf{X}_M\theta)$ and $\overline{\pi}^\theta = \overline{\mathbb{E}}_\theta[Y]$.

Let us now introduce what we will call the C-cube conditions in this chapter.

- $\mathbf{C}_1$ We are able to sample from the distribution $\overline{\mathbb{P}}_\theta$.
- $\mathbf{C}_2$ Under appropriate conditions, we have the following CLT

  $$u^\top [\overline{V}_N(\theta^*)]^{1/2} (\widehat{\theta} - \overline{\theta}) \xrightarrow[N \to +\infty]{(d)} \mathcal{N}(0,1),$$

  where u is a unit s-vector (with $s = |M|$), $\widehat{\theta} = \Psi(\mathbf{X}_M^\top Y)$ is the MLE, and $\overline{V}_N(\theta^*)$ is a positive semi-definite $(s \times s)$-matrix.
- $\mathbf{C}_3$ We are able to compute efficiently $\Psi(\rho)$ for any $\rho \in \mathbb{R}^s$.

The C-cube conditions refer to $\mathbf{C}_1$: Conditional Sampling, $\mathbf{C}_2$: Conditional CLT and $\mathbf{C}_3$: Computation of Ψ. In any GLM where the C-cube conditions are satisfied, one can adapt the methods of this chapter to design asymptotically valid PSI procedures with respect to the selected model.

6.3.3 Discussion

Duality between conditional MLE and debiasing approaches. Oversimplifying the situation, our approach could be understood as the dual counterpart of the one from Taylor and Tibshirani [2018] in the sense that the former paper is first focused on getting an (unconditional) CLT and deal with the selection event in a second phase. On the contrary, we are first focused on the conditional distribution (*i.e.,* we want to be able to sample from the conditional distribution) while the asymptotic (conditional) distribution considerations come thereafter.

What about the saturated model? In this section, we have presented the conditional MLE approach in the selected model. Nevertheless, we provide in this chapter asymptotically valid post-selection inference procedures on $\mathbf{X}_M^\top \pi^*$ in the saturated model for logistic regression. Let us stress that this approach could also be adapted to obtain analogous methods in other GLMs.

Comprehensive comparison between our work and the one from Taylor and Tibshirani [2018]. In Taylor and Tibshirani [2018], the authors consider only the more restrictive framework of the selected model where $\mathbf{X}\vartheta^* = \mathbf{X}_M\theta^*$ for some $\theta^* \in \mathbb{R}^s$. Their method allows to conduct PSI inference on any linear transformation of θ^* (including in particular the local coordinates θ_j^* for $j \in [s]$), and can be efficiently used in practice. The authors do not provide a formal proof of their claim but rather motivate their approach with asymptotic arguments where they consider in particular that $\vartheta^* = N^{-1/2}\beta^*$ where each entry of β^* is independent of N.
On the other hand, we present *global* PSI methods in both the saturated and the selected models, in the sense that statistical inference is conducted on the vector-valued parameter of interest. Our methods are computationally more expensive than the one from Taylor and Tibshirani [2018], but they are proved to be asymptotically valid under some set of assumptions that we discuss in details in Section 6.5.4. Table 6.7 sums up this comparison.

	Taylor and Tibshirani [2018]	Our work
Selected model	✓	✓
Saturated model	✗	✓
Hypotheses tested in the selected model	Simple: $\theta_j^* = [\theta_0^*]_j$ for some j	Multiple: $\theta^* = \theta_0^*$
Formal proof	✗	✓
Assumption on $\vartheta^* = \alpha_N^{-1}\beta^*$ with entries of β^* independent of N	For the theoretical sketches supporting their result, they consider $\alpha_N = N^{1/2}$.	Require $\alpha_N = \omega(N^{1/2})$, that could be weakened (Sec.6.5.4).
Low computational cost	✓	✗

Table 6.7: Comparison between our work and the one from Taylor and Tibshirani [2018].

The logistic regression. In the remaining sections of this chapter, we focus on the logistic regression case. This means in particular that $\xi(x) = \ln(1 + \exp(x))$ is the softmax function and its derivative $\sigma(x) = (1 + \exp(-x))^{-1}$ is the sigmoid function. We prove that the C-cube conditions hold in this framework under some assumptions, and we describe our methods for PSI for both the selected and the saturated models.
Note that our work should be understood as an extension of the one from Meir and Drton [2017] to the SLR. Indeed, the authors of the former paper propose a method to compute the conditional MLE after model selection in the linear model. They show empirically that the proposed confidence intervals are close to the desired level but they are not able to provide theoretical justification of their approach.

6.4 Sampling from the conditional distribution

In this section, we present an algorithm based on a simulated annealing approach that is proved to sample states $Y^{(t)}$ uniformly distributed on the selection event E_M for any $M \subseteq [d]$ with cardinality s in the asymptotic regimes as $t \to \infty$. From this section, we consider the case of the logistic regression where we recall that $Y = (y_i)_{i\in[N]}$ and for all $i \in [N]$, $y_i \sim \mathrm{Ber}(\pi_i^*)$ with $\pi^* = \sigma(\mathbf{X}\vartheta^*)$.

6.4.1 Numerical method to approximate the selection event

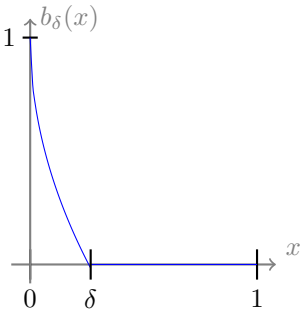
In this section, we present a simulated annealing algorithm to approximate the selection event E_M for some set $M \subseteq [d]$. From Proposition 6.1 and the KKT conditions from (6.13), we know that the selection

event E_M can be written as

$$E_M = \left\{ Y \in \{0,1\}^N \mid \mathbb{1}_{\|\widehat{S}_{-M}(Y)\|_\infty - 1 < 0},\ \mathbb{1}_{1=\min_{k\in M}\{|\widehat{S}_k(Y)|\}} \right\}. \tag{6.31}$$

Based on the expression of E_M given in Eq.(6.31), we introduce the function

$$b_\delta(x) = 1 - \sqrt{\left(\frac{x}{\delta}\right) \wedge 1}\,,$$



for some $\delta > 0$ and we define the energy

$$\mathcal{E}(Y) := \max\left\{p_1(Y)\ ,\ p_2(Y)\right\},$$

where

$$p_1(Y) := b_\delta\left(1 - \|\widehat{S}_{-M}(Y)\|_\infty\right) \quad \text{and} \quad p_2(Y) := \frac{1}{|M|}\sum_{k\in M}(1 - |\widehat{S}_k(Y)|).$$

The energy $\mathcal{E}$ measures how close some vector $Y \in \{0,1\}^N$ is to E_M. With Lemma 6.5, we make this claim rigorous by proving that for $\delta > 0$ small enough, the selection event E_M corresponds to the set of vectors $Y \in \{0,1\}^N$ satisfying $\mathcal{E}(Y) = 0$.

Lemma 6.5. *For any $M \subseteq [d]$, there exists $\delta_c := \delta_c(M, \mathbf{X}, \lambda) > 0$ such that for all $\delta \in (0, \delta_c)$, the selection event $E_M = \{Y \in \{0,1\}^N \mid \widehat{M}(Y) = M\}$ is equal to the set*

$$\left\{Y \in \{0,1\}^N \mid p_1(Y) = 0 \quad \textit{and} \quad p_2(Y) = 0\right\}.$$

Proof. Let us consider some $\delta \in (0, \delta_c)$ where

$$\delta_c := \min_{Y\in E_M}\{1 - \|\widehat{S}_{-M}(Y)\|_\infty\}.$$

Note that Eq.(6.31) ensures that for any $Y \in E_M$, $\|\widehat{S}_{-M}(Y)\|_\infty < 1$. This implies that $\delta_c > 0$ since the set E_M is finite.
It is obvious that for any $Y \in \{0,1\}^N$, the fact that $p_2(Y) = 0$ is equivalent to $\min_{k\in M}|\widehat{S}_k(Y)| = 1$. Moreover, thanks to our choice for the constant δ, it also holds that $p_1(Y) = 0$ is equivalent to $\|\widehat{S}_{-M}(Y)\|_\infty < 1$. The characterization of the selection event E_M given by Eq.(6.31) allows to conclude the proof. □

Lemma 6.5 states that-provided δ is small enough–the selection event E_M corresponds to the set of global minimizers of the energy $\mathcal{E} : \{0,1\}^N \to \mathbb{R}_+$. This characterization allows us to formulate a simulating annealing (SA) procedure in order to estimate E_M. Let us briefly recall that simulated annealing algorithms are used to estimate the set of global minimizers of a given function. At each time step, the algorithm considers some neighbour of the current state and probabilistically decides between moving to the proposed neighbour or staying at its current location. While a transition to a state inducing a lower energy compared to the current one is always performed, the probability of transition towards a neighbour that leads to increase the energy is decreasing over time. The precise expression of the probability of transition is driven by a chosen *cooling schedule* $(\mathrm{T}_t)_t$ where T_t are called *temperatures* and vanish as $t \to \infty$. Intuitively, in the first iterations of the algorithm the temperature is high and we are likely to accept most of the transitions proposed by the SA. In that way, we give our algorithm the chance to escape from local minima. As time goes along, the temperature decreases and we expect to end up at a global minimum of the function of interest.
We refer to [Brémaud, 2013, Chapter 12] for further details on SA. Our method is described in Algorithm 5 and in the next section, we provide theoretical guarantees. In Algorithm 5, $P : \{0,1\}^N \times \{0,1\}^N \to [0,1]$ is the Markov transition kernel such that for any $Y \in \{0,1\}^N$, $P(Y,\cdot)$ is the probability measure on $\{0,1\}^N$ corresponding to the uniform distribution on the vectors in $\{0,1\}^N$ that differs from Y in exactly one coordinate.

Algorithm 5 SEI-SLR: Selection Event Identification for SLR

Data: $\mathbf{X}$, Y, λ, K_0, T

Compute $\hat{\vartheta}^\lambda \in \underset{\vartheta\in\mathbb{R}^d}{\arg\min}\{\mathcal{L}_N(\vartheta,(Y,\mathbf{X}))+\lambda\|\vartheta\|_1\}$
Set $M=\{k\in[d]\mid \hat{\vartheta}^\lambda_k\neq 0\}$
$Y^{(0)}\leftarrow Y$
for $t=1$ to T **do**
  $Y^{\mathrm{c}}\sim P(Y^{(t-1)},\cdot)$
  $\hat{\vartheta}^{\lambda,\mathrm{c}}\in \underset{\vartheta\in\mathbb{R}^d}{\arg\min}\{\mathcal{L}_N(\vartheta,(Y^{\mathrm{c}},\mathbf{X}))+\lambda\|\vartheta\|_1\}$
  $\widehat{S}(Y^{\mathrm{c}})=\frac{1}{\lambda}\mathbf{X}^\top(Y^{\mathrm{c}}-\sigma(\mathbf{X}\hat{\vartheta}^{\lambda,\mathrm{c}}))$
  $\Delta\mathcal{E}=\mathcal{E}(Y^{\mathrm{c}})-\mathcal{E}(Y^{(t-1)})$
  $U\sim\mathcal{U}([0,1])$
  $\mathrm{T}_t\leftarrow\frac{K_0}{\log(t+1)}$
  if $\exp\left(-\frac{\Delta\mathcal{E}}{\mathrm{T}_t}\right)\geq U$ **then**
    $Y^{(t)}\leftarrow Y^{\mathrm{c}}$
  end if
end for

6.4.2 Proof of convergence of the algorithm

To provide theoretical guarantees on our methods in the upcoming sections, we need to understand what is the distribution of $Y^{(t)}$ as $t\to\infty$. This is the purpose of Proposition 6.6 which shows that the SEI-SLR algorithm generates states uniformly distributed on E_M in the asymptotic $t\to\infty$.

Proposition 6.6. *[Brémaud, 2013, Example 12.2.12]*
For a cooling schedule satisfying $\mathrm{T}_t\geq 2^{N+1}/\log(t+1)$, *the limiting distribution of the random vectors* $Y^{(t)}$ *is the uniform distribution on the selection event* E_M.

Proposition 6.6 has the important consequence that we are able to compute the distribution of the binary vector $Y=(y_i)_{i\in[N]}$ where each y_i is a Bernoulli random variable with parameter $\pi_i^*\in(0,1)$ conditional on the selection event. The formal presentation of this result is given by Proposition 6.7 which will be the cornerstone of our inference procedures presented in Section 6.5.

Proposition 6.7. *Let us consider* $M\subseteq[d]$ *and some* $\vartheta^*\in\mathbb{R}^d$. *We consider the random vector* Y *with distribution* $\overline{\mathbb{P}}_{\pi^*}$ *where* $\pi^*=\sigma(\mathbf{X}\vartheta^*)$. *For a cooling schedule satisfying* $\mathrm{T}_t\geq 2^{N+1}/\log(t+1)$, *it holds for any function* $h:\{0,1\}^N\to\mathbb{R}$,

$$\frac{\sum_{t=1}^T h(Y^{(t)})\mathbb{P}_{\pi^*}(Y^{(t)})}{\sum_{t=1}^T \mathbb{P}_{\pi^*}(Y^{(t)})}\underset{T\to\infty}{\to}\overline{\mathbb{E}}_{\pi^*}\left[h(Y)\right]\quad \textit{almost surely}.$$

Proof. Let us consider some map $h:\{0,1\}^N\to\mathbb{R}$. Then,

$$\overline{\mathbb{E}}_{\pi^*}\left[h(Y)\right]=\frac{\sum_{y\in E_M}h(y)\mathbb{P}_{\pi^*}(y)}{\sum_{y\in E_M}\mathbb{P}_{\pi^*}(y)}=\frac{\mathbb{E}(h(U_M)\mathbb{P}_{\pi^*}(Y=U_M))}{\mathbb{E}(\mathbb{P}_{\pi^*}(Y=U_M))},$$

where U_M is a random variable taking values in $\{0,1\}^N$ which is uniformly distributed over E_M. Then the conclusion directly follows from Proposition 6.6. □

6.5 Conditional Central Limit Theorems

6.5.1 Preliminaries

Before presenting our conditional CLTs, let us remind the framework in which we state our asymptotic results. Let $(d_N)_{N\in\mathbb{N}}$ be a non-decreasing sequence of positive integers converging to $d_\infty\in\mathbb{N}\cup\{+\infty\}$ and let $s\in[d_1,d_\infty]\cap\mathbb{N}$. For any N, we consider $[\vartheta^*]^{(N)}\in\mathbb{R}^{d_N}$, $\lambda^{(N)}>0$, $M^{(N)}\subseteq[d_N]$ with cardinality s and a design matrix $\mathbf{X}^{(N)}\in\mathbb{R}^{N\times d_N}$. We recall the definitions of the selection event $E_M^{(N)}$ corresponding to the tuple $(\lambda^{(N)},M^{(N)},\mathbf{X}^{(N)})$ and of the conditional probability distribution $\overline{\mathbb{P}}_{\pi^*}^{(N)}$ given in Section 6.2.4. We assume that it holds

- $K := \sup_{N\in\mathbb{N}} \max_{i\in[N], j\in M^{(N)}} |\mathbf{X}^{(N)}_{i,j}| < \infty$,
- there exist constants $C, c > 0$ (independent of N) such that for any $N \in \mathbb{N}$,

$$cN \leq \lambda_{\min}\big([\mathbf{X}^{(N)}_{M^{(N)}}]^\top \mathbf{X}^{(N)}_{M^{(N)}}\big) \leq \lambda_{\max}\big([\mathbf{X}^{(N)}_{M^{(N)}}]^\top \mathbf{X}^{(N)}_{M^{(N)}}\big) \leq CN.$$

Remark. Note that the latter assumption holds in particular if the matrices $\left(\frac{\mathbf{X}^{(N)}}{\sqrt{N}}\right)_{N\in\mathbb{N}}$ satisfy (uniformly) the so-called s-Restricted Isometry Property (RIP) condition [cf. Wainwright, 2019, Definition 7.10]. Let us recall that a matrix $A \in \mathbb{R}^{N\times p}$ satisfies the s-RIP condition if there exists a constant $\delta_s \in (0,1)$ such that for any $N \times s$ submatrix A_s of A, it holds

$$1 - \delta_s \leq \lambda_{\min}(A_s^\top A_s) \leq \lambda_{\max}(A_s^\top A_s) \leq 1 + \delta_s.$$

In Section 6.5.2, we start by presenting our first CLT for $[\mathbf{X}^{(N)}_M]^\top Y$ where Y is distributed according to $\overline{\mathbb{P}}^{(N)}_{\pi^*}$. This will be the cornerstone of our PSI procedures holding for the saturated model and presented in Section 6.6.2. Thereafter, we prove in Section 6.5.3 a CLT for the conditional unpenalized MLE $\widehat{\theta}$ working with the design $\mathbf{X}^{(N)}_M$ (see Eq.(6.28)). This conditional CLT will be the key theoretical ingredient to derive the PSI methods presented in Section 6.6.1 when considering the selected model.

The proofs of our conditional CLTs make use of [Bardet et al., 2008, Thm.1] and rely on triangular arrays $\vec{\xi} := \big((\xi_{i,N})_{i\in[N]},\, N \in \mathbb{N}\big)$ where $\xi_{i,N}$ is a random vector in $\mathbb{R}^s$ and is a function of the deterministic quantities $\lambda^{(N)}$, $\mathbf{X}^{(N)}$, $M^{(N)}$ and of the random variable Y with probability distribution $\overline{\mathbb{P}}^{(N)}_{\pi^*}$. Most dependent CLTs have been proven for causal time series (typically satisfying some mixing condition) and are not well-suited to our case since conditioning on the selection event introduces a complex dependence structure.

$$\begin{array}{cccccc}
\xi_{1,1} & & & & & \\
\xi_{1,2} & \xi_{2,2} & & & & \\
\xi_{1,3} & \xi_{2,3} & \xi_{3,3} & & & \\
\dots & \dots & \dots & \dots & & \\
\xi_{1,N} & \xi_{2,N} & \xi_{3,N} & \dots & \xi_{N,N} & \\
\dots & \dots & \dots & \dots & \dots & \dots
\end{array}$$

The dependent Lindeberg CLT from [Bardet et al., 2008, Thm.1] gives us the opportunity to find conditions involving mainly the covariance matrix of Y under which our conditional CLTs hold. More precisely, we provide conditions ensuring that the lines of the $\mathbb{R}^s$-valued process indexed by a triangular system $\vec{\xi}$ satisfy some Lindeberg's condition. Let us stress that we discuss the assumptions of the theorems presented in Sections 6.5.2 and 6.5.3 in Section 6.5.4.

To alleviate this notational burden, we will not specify the dependence on N in the remainder of the chapter, meaning that we will simply refer to $\mathbf{X}^{(N)}$, $M^{(N)}$, d_N, $[\vartheta^*]^{(N)}$, $\overline{\mathbb{P}}^{(N)}_{\pi^*}, \dots$ as $\mathbf{X}$, M, d, ϑ^*, $\overline{\mathbb{P}}_{\pi^*}, \dots$. Nevertheless, let us stress again that the integer s is fixed and does not depend on N in our work.

6.5.2 A conditional CLT for the saturated model

We aim at providing a multiple testing procedure and a confidence interval for the parameter π^* conditionally to the selection event E_M. To do so, we prove in this section a CLT for $\mathbf{X}_M^\top Y$ when Y is a random variable on $\{0,1\}^N$ following the multivariate Bernoulli distribution with parameter $\pi^* \in [0,1]^N$ conditional on the event $\{Y \in E_M\}$. Let us first recall the notation for the distribution of Y conditional on E_M in the saturated model

$$\overline{\mathbb{P}}_{\pi^*}(Y) \propto \mathbb{1}_{E_M}(Y)\mathbb{P}_{\pi^*}(Y),$$

where the symbol $\propto$ means "proportional to". In the following, we will denote by $\overline{\mathbb{E}}_{\pi^*}$ the expectation with respect to $\overline{\mathbb{P}}_{\pi^*}$. With Theorem 6.8, we give a conditional CLT that holds under some conditions that involve in particular the covariance matrix of the response Y under the distribution $\overline{\mathbb{P}}_{\pi^*}$, namely

$$\overline{\Gamma}^{\pi^*} := \overline{\mathbb{E}}_{\pi^*}\left[(Y - \overline{\pi}^{\pi^*})(Y - \overline{\pi}^{\pi^*})^\top\right] \in [-1,1]^{N\times N},$$

where $\overline{\pi}^{\pi^*} = \overline{\mathbb{E}}_{\pi^*}[Y]$.

Theorem 6.8. *We keep the notations and assumptions from Section 6.5.1. We denote $\pi^* = \sigma(\mathbf{X}\vartheta^*)$ and Y the random vector taking values in $\{0,1\}^N$ and distributed according to $\overline{\mathbb{P}}_{\pi^*}$. Assume further that*

1. $\sum_{i=1}^{N} \sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]} \mathbf{X}_{[i-1],M}\|_F \left(1 - 2\overline{\pi}_i^{\pi^*}\right)^2} \underset{N\to+\infty}{=} o(N)$,
2. *there exists* $\overline{\sigma}^2_{\min} > 0$ *such that for any* $N \in \mathbb{N}$, $\overline{\pi}_i^{\pi^*}(1 - \overline{\pi}_i^{\pi^*}) \geq \overline{\sigma}^2_{\min}$ *for all* $i \in [N]$.

Then it holds

$$u^\top [\overline{G}_N(\pi^*)]^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*}) \underset{N\to+\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}(0,1),$$

where u *is a unit* s*-vector and where* $\overline{G}_N(\pi^*) := \mathbf{X}_M^\top \mathrm{Diag}((\overline{\sigma}^{\pi^*})^2)\mathbf{X}_M$ *with* $(\overline{\sigma}^{\pi^*})^2 := \overline{\pi}^{\pi^*} \odot (1 - \overline{\pi}^{\pi^*})$.

6.5.3 A conditional CLT for the selected model

We now work under the condition that there exists $\theta^* \in \mathbb{R}^s$ such that $\mathbf{X}_M \theta^* = \mathbf{X}\vartheta^*$. Given some $Y \in \{0,1\}^N$ and provided that $\mathbf{X}_M^\top Y \in \mathrm{Im}(\Xi)$, $\Psi(\mathbf{X}_M^\top Y)$ is the MLE $\widehat{\theta}$ of the unpenalized logistic model. [Sur and Candès, 2019, Theorem 1] ensures that the MLE exists asymptotically almost surely when Y is distributed as $\mathbb{P}_{\theta^*}$. When the distribution of Y is $\overline{\mathbb{P}}_{\theta^*}$, we prove in Section 6.8.5 a weaker counterpart of this result showing that for N large enough, the MLE exists with high probability.
We aim at providing a multiple testing procedure and a confidence interval for the parameter θ^* conditionally on the selection event. To do so, we first prove a CLT for the MLE $\widehat{\theta}$ when Y is distributed according to $\overline{\mathbb{P}}_{\theta^*}$ (*i.e.,* Y is a random variable on $\{0,1\}^N$ following the multivariate Bernoulli distribution with parameter $\sigma(\mathbf{X}_M\theta^*)$ conditioned on the event $\{Y \in E_M\}$). The unconditional MLE $\widehat{\theta}$ (using only the features indexed by M) is known to be consistent and asymptotically efficient meaning that when Y is distributed according to $\mathbb{P}_{\theta^*}$,

$$u^\top [G_N(\theta^*)]^{1/2} (\widehat{\theta} - \theta^*) \underset{N\to+\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}(0,1), \tag{6.32}$$

where u is a unit s-vector and where

$$G_N(\theta) := \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M = \mathbf{X}_M^\top \mathrm{Diag}((\sigma^\theta)^2)\mathbf{X}_M,$$

is the Fisher information matrix with $(\sigma^\theta)^2 := \pi^\theta \odot (1 - \pi^\theta)$ and $\pi^\theta = \mathbb{E}_\theta[Y]$.
In the following, we will consider the natural counterpart of the Fisher information matrix $G_N(\theta^*)$ when we work under the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$,

$$\overline{G}_N(\theta^*) := \mathbf{X}_M^\top \mathrm{Diag}((\overline{\sigma}^{\theta^*})^2)\mathbf{X}_M, \quad (\overline{\sigma}^{\theta^*})^2 := \overline{\pi}^{\theta^*} \odot (1 - \overline{\pi}^{\theta^*}),\ \overline{\pi}^{\theta^*} = \overline{\mathbb{E}}_{\theta^*}[Y].$$

Theorem 6.9 proves that the MLE $\widehat{\theta}$ under the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ also satisfies a CLT analogous to Eq.(6.32) by replacing respectively θ^* and $G_N(\theta^*)^{1/2}$ by $\overline{\theta}(\theta^*)$ (cf. Eq.(6.29)) and $[\overline{G}_N(\theta^*)]^{-1/2} G_N(\overline{\theta}(\theta^*))$. This conditional CLT holds under some conditions that involve in particular the covariance matrix of the response Y under the distribution $\overline{\mathbb{P}}_{\theta^*}$, namely

$$\overline{\Gamma}^{\theta^*} = \overline{\mathbb{E}}_{\theta^*}\left[(Y - \overline{\pi}^{\theta^*})(Y - \overline{\pi}^{\theta^*})^\top\right] \in [-1,1]^{N\times N}.$$

Theorem 6.9. *We keep the notations and assumptions from Section 6.5.1. Let us consider* $\theta^* \in \mathbb{R}^s$ *and let us denote by* Y *the random vector taking values in* $\{0,1\}^N$ *and distributed according to* $\overline{\mathbb{P}}_{\theta^*}$. *Assume further that*

1. $\sum_{i=1}^{N} \sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\theta^*}_{[i-1],[i-1]} \mathbf{X}_{[i-1],M}\|_F \left(1 - 2\overline{\pi}_i^{\theta^*}\right)^2} \underset{N\to+\infty}{=} o(N)$,
2. *there exists* $\overline{\sigma}^2_{\min} > 0$ *such that for any* N *and for any* $i \in [N]$,
$$\overline{\pi}_i^{\theta^*}(1 - \overline{\pi}_i^{\theta^*}) \wedge \sigma'(\mathbf{X}_{i,M}\overline{\theta}(\theta^*)) \geq \overline{\sigma}^2_{\min}.$$
3. *there exists some* $\mathfrak{K} > 0$ *such that for any* $N \in \mathbb{N}$,
$$\mathrm{Tr}\left[\overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M\right] < \mathfrak{K}.$$

Then,

$$u^\top [\overline{G}_N(\theta^*)]^{-1/2} G_N(\overline{\theta}(\theta^*))\big(\widehat{\theta} - \overline{\theta}(\theta^*)\big) \underset{N\to+\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}(0,1),$$

where u is a unit s-vector and where we recall that $\widehat{\theta} = \Psi(\mathbf{X}_M^\top Y)$ is the MLE.

The proof of Theorem 6.9 can be found with full details in Section 6.8.5 and we only provide here the main arguments. First we use Theorem 6.8 that shows that the distribution of $[\overline{G}_N(\theta^*)]^{-1/2} L_N(\overline{\theta}, Z^M)$ is asymptotically Gaussian using a Lindeberg Central Limit Theorem for dependent random variables from Bardet et al. [2008]. Then, we show that for N large enough, the following holds with high probability: the MLE $\widehat{\theta}$ exists and is contained within an ellipsoid centered at $\overline{\theta}$ with vanishing volume. This kind of result has already been studied in Liang and Du [2012] but the proof provided by Liang and Du is wrong (Eq.(3.7) is in particular not true). As far as we know, we are the first to provide a correction of this proof in Section 6.8.5. Let us also stress that working with the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ brings extra-technicalities that need to be handled carefully.
Using this consistency of $\widehat{\theta}$ together with the smoothness of the map $\theta \mapsto L_N(\theta, Z^M)$, one can convert the previously established result for $[\overline{G}_N(\theta^*)]^{-1/2} L_N(\overline{\theta}, Z^M) = [\overline{G}_N(\theta^*)]^{-1/2}(L_N(\overline{\theta}, Z^M) - L_N(\widehat{\theta}, Z^M))$ into a CLT for $\widehat{\theta}$.

6.5.4 Discussion

In this section, we discuss *informally* the assumptions of both Theorems 6.8 and 6.9. The conditions of Theorems 6.8 and 6.9 can be seen at first glance as arcane or restrictive. Without pretending that those conditions are easy to check in practice, looking at these requirements through the lens of the usual asymptotic alternative where ϑ^* itself depends on N gives a different perspective. Such assumption on ϑ^* has been considered for example in Bunea [2008] or [Taylor and Tibshirani, 2018, Section 3.1]. Following this line of work, we consider that $\vartheta^* = \alpha_N^{-1}\beta^*$ where each entry of β^* is independent of N and $(\alpha_N)_N$ is a sequence of increasing positive numbers such that $\alpha_N \underset{N\to\infty}{\to} +\infty$. We further assume β^* is s^*-sparse with support M^* (and with s^* independent of N). Let us analyze the conditions of our theorems in this framework by considering that $E_M = \{0,1\}^N$ (i.e. there is no conditioning). Then, condition 3 of Theorem 6.9 holds automatically since in this case $\mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M = G_N(\theta^*)$ and $\overline{G}_N^{-1} = [G_N(\theta^*)]^{-1}$, meaning that $\mathfrak{K} = s$ works. The condition 2 of Theorems 6.8 and 6.9 holds also automatically since $\alpha_N \underset{N\to\infty}{\to} +\infty$, while the condition 1 is satisfied as soon as $\alpha_N \underset{N\to\infty}{=} \omega(N^{1/2})$.

The quantity α_N is quantifying the dependence arising from conditioning on the selection event: the weaker the dependence between the entries of the random response $Y \sim \overline{\mathbb{P}}_{\pi^*}$, the smaller α_N can be chosen while preserving the asymptotic normal distribution. Note that in the papers Bunea [2008] and [Taylor and Tibshirani, 2018, Section 3.1], the authors typically consider the case where $\alpha_N \underset{N\to\infty}{\sim} N^{1/2}$, corresponding to the regime at which the validity of our CLTs may be questioned based on the simple analysis previously conducted. Nevertheless, we stress that stronger assumptions on the design could allow to bypass this apparent limitation. A promising line of investigation is the following: taking a closer at the proofs of Theorems 6.8 and 6.9, one can notice that the condition 1 can actually be weakened by

$$\min_{\nu\in\mathfrak{S}_N} \sum_{i=1}^N \sqrt{\|(\mathbf{X}_{\nu([i-1]),M})^\top \overline{\Gamma}^{\pi^*}_{\nu([i-1]),\nu([i-1])} \mathbf{X}_{\nu([i-1]),M}\|_F \left(1 - 2\overline{\pi}^{\pi^*}_{\nu(i)}\right)^2} \underset{N\to+\infty}{=} o(N),$$

where $\mathfrak{S}_N$ is the set of permutations of $[N]$.

6.6 Selective inference

6.6.1 In the selected model

6.6.1.1 Multiple testing procedure

We keep the notations and the assumptions of Theorem 6.9. Given some $\theta_0^* \in \mathbb{R}^s$, we consider the hypothesis test with null and alternative hypotheses defined by

$$\mathbb{H}_0 \ : \ \{\theta^* = \theta_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\theta^* \neq \theta_0^*\}. \tag{6.33}$$

The CLT from Theorem 6.9 naturally leads us to introduce the ellipsoid W_N given by

$$W_N := \left\{ Y \in \{0,1\}^N \;\middle|\; \begin{array}{l} \diamond\ \mathbf{X}_M^\top Y \in \mathrm{Im}(\Xi) \\ \diamond\ \left\| [\overline{G}_N(\theta_0^*)]^{-1/2} G_N(\overline{\theta}(\theta_0^*)) \left(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta_0^*) \right) \right\|_2^2 > \chi^2_{s,1-\alpha} \end{array} \right\},$$

where $\chi^2_{s,1-\alpha}$ is the quantile of order $1-\alpha$ of the χ^2 distribution with s degrees of freedom. If $\overline{\pi}^{\theta_0^*}$ was known, we could compute $\overline{\theta}(\theta_0^*)$ (using Eq.(6.30)) and thus $\overline{G}_N(\theta_0^*)$. Then the test with rejection region W_N would be asymptotically of level α since Theorem 6.9 gives that

$$\overline{\mathbb{P}}_{\theta_0^*}\left(Y \in W_N\right) \underset{N \to +\infty}{\longrightarrow} \alpha.$$

Based on this result, we construct an asymptotically valid multiple testing procedure for the test (6.33). Our method consists in finding an estimate of the parameter $\overline{\pi}^{\theta_0^*}$ in order to approximate the rejection region W_N with a Monte-Carlo approach. From Proposition 6.6, we know that under an appropriate cooling scheme, the asymptotic distribution of the states visited by our SEI-SLR algorithm (cf. Algorithm 5) is the uniform distribution on the selection event. We deduce that under the null, we are able to estimate $\overline{\pi}^{\theta^*}$ and thus $\overline{\theta}$ using Eq.(6.30). This leads to the testing procedure presented in Proposition 6.10, whose proof is postponed to Section 6.8.7.

Proposition 6.10. *We keep notations and assumptions of Theorem 6.9. We consider two independent sequences of vectors $(Y^{(t)})_{t\geq 1}$ and $(Z^{(t)})_{t\geq 1}$ generated by Algorithm 5. Let us denote*

$$\widetilde{\pi}^{\theta_0^*} = \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)}) Y^{(t)}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})}, \quad \widetilde{\theta} = \Psi(\mathbf{X}_M^\top \widetilde{\pi}^{\theta_0^*}), \quad \widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}\left(\widetilde{\pi}^{\theta_0^*} \odot (1 - \widetilde{\pi}^{\theta_0^*}) \right) \mathbf{X}_M,$$

$$\textit{and } W_N := \left\{ Y \in \{0,1\}^N \;\middle|\; \begin{array}{l} \diamond\ \mathbf{X}_M^\top Y \in \mathrm{Im}(\Xi) \\ \diamond\ \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Y) - \widetilde{\theta} \right) \right\|_2^2 > \chi^2_{s,1-\alpha} \end{array} \right\}.$$

Then the procedure consisting in rejecting the null hypothesis $\mathbb{H}_0$ when

$$\zeta_{N,T} := \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)}) \mathbb{1}_{Z^{(t)} \in \widetilde{W}_N}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)})} > \alpha,$$

has an asymptotic level lower than α in the sense that for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any $N \geq N_0$ it holds,

$$\mathbb{P}\Big(\bigcup_{T_N \in \mathbb{N}} \bigcap_{T \geq T_N} \{\zeta_{N,T} \leq \alpha + \epsilon\} \Big) = 1.$$

6.6.1.2 Asymptotic confidence region

In the previous section, we proved that the MLE $\widehat{\theta}$ satisfies a CLT with a centering vector that is not the parameter of interest θ^*. Two questions arises at this point.

1. How can we compute a relevant estimate for θ^*?
2. Can we provide theoretical guarantees regarding this estimate?

Proposition 6.11 answers both questions. It provides a valid confidence region with asymptotic level $1-\alpha$ for any estimate $\theta^\bigstar$ of θ^* where the width of the confidence region is asymptotically driven by $\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2$. The proof of Proposition 6.11 can be found in Section 6.8.8.

Proposition 6.11. *We keep notations and assumptions of Theorem 6.9 and we assume further that there exist $p \in [1,\infty]$ and $\kappa, R > 0$ such that*

$$\theta^* \in \mathbb{B}_p(0,R) \quad \textit{and} \quad \forall \theta \in \mathbb{B}_p(0,R), \quad \lambda_{\min}(\overline{\Gamma}^\theta) \geq \kappa,$$

where $\mathbb{B}_p(0,R) := \{\theta \in \mathbb{R}^s \mid \|\theta\|_p \leq R\}$. Let us consider any estimator $\theta^\bigstar \in \mathbb{B}_p(0,R)$ of θ^*. Then the probability of the event

$$\|\theta^* - \theta^\bigstar\|_2 \leq C\,(\kappa c)^{-1} \left\{ \|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2 + \|(\sigma^{\overline{\theta}})^{-2}\|_\infty \left(Nc^2/C\right)^{-1/2} \sqrt{\chi^2_{s,1-\alpha}} \right\},$$

tends to $1-\alpha$ *as* $N \to \infty$. *We recall that* $(\sigma^{\overline{\theta}})^2 = \sigma'(\mathbf{X}_M \overline{\theta}(\theta^*))$.

Remarks. In Proposition 6.11, note that the constants c and C can be easily computed from the design matrix. Nevertheless, we point out that the confidence region from Proposition 6.11 involves two constants (namely κ and $\sigma^{\overline{\theta}}$) that cannot be *a priori* easily computed in practice.
Proposition 6.11 proves that when N is large enough, the size of our confidence region is driven by the distance $\|\overline{\theta}(\theta^\star) - \widehat{\theta}\|_2$. This remark motivates us to choose $\theta^\star$ among the minimizers of the function

$$m : \theta \mapsto \|\overline{\theta}(\theta) - \widehat{\theta}\|_2^2.$$

In the sake of minimizing m, a large set of methods are at our disposal. In Section 6.7, we propose a deep learning and a gradient descent approach for our numerical experiments.

6.6.2 In the saturated model

6.6.2.1 Multiple testing procedure

We keep the notations and the assumptions of Theorem 6.8. Given some $\pi_0^* \in \mathbb{R}^N$, we consider the hypothesis test with null and alternative hypotheses defined by

$$\mathbb{H}_0 \ : \ \{\pi^* = \pi_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\pi^* \neq \pi_0^*\}. \tag{6.34}$$

The CLT from Theorem 6.8 naturally leads us to introduce the ellipsoid W_N given by

$$W_N = \left\{ Y \in \{0,1\}^N \mid \left\| [\overline{G}_N(\pi_0^*)]^{-1/2} \mathbf{X}_M^\top \left(Y - \overline{\pi}^{\pi_0^*} \right) \right\|_2^2 \geq \chi^2_{s,1-\alpha} \right\},$$

where $\chi^2_{s,1-\alpha}$ is the quantile of order $1-\alpha$ of the χ^2 distribution with s degrees of freedom. If $\overline{\pi}^{\pi_0^*}$ was known, we could compute $\overline{G}_N(\pi_0^*)$. Then the test with rejection region W_N would be asymptotically of level α since Theorem 6.8 gives that

$$\overline{\mathbb{P}}_{\pi_0^*} \left(Y \in W_N \right) \underset{N \to +\infty}{\longrightarrow} \alpha.$$

Based on this result, we construct an asymptotically valid multiple testing procedure for the test (6.34). Our method consists in finding an estimate of the parameter $\overline{\pi}^{\pi_0^*}$ in order to approximate the rejection region W_N with a Monte-Carlo approach. From Proposition 6.6, we know that under an appropriate cooling scheme, the asymptotic distribution of the states visited by the SEI-SLR algorithm (cf. Algorithm 5) is the uniform distribution on the selection event. We deduce that under the null, we are able to estimate $\overline{\pi}^{\pi_0^*}$ and thus $\overline{G}_N(\pi_0^*)$. This leads to the testing procedure presented in Proposition 6.12, whose proof is strictly analogous to the one of Proposition 6.10.

Proposition 6.12. *We keep notations and assumptions of Theorem 6.8. We consider two independent sequences of vectors* $(Y^{(t)})_{t\geq 1}$ *and* $(Z^{(t)})_{t\geq 1}$ *generated by Algorithm 5. Let us denote*

$$\widetilde{\pi}^{\pi_0^*} = \frac{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Y^{(t)}) Y^{(t)}}{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Y^{(t)})}, \qquad \widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}\left(\widetilde{\pi}^{\pi_0^*} \odot (1 - \widetilde{\pi}^{\pi_0^*}) \right) \mathbf{X}_M,$$

and $\widetilde{W}_N := \left\{ Y \in \{0,1\}^N \mid \left\| \widetilde{G}_N^{-1/2} \mathbf{X}_M^\top \left(Y - \widetilde{\pi}^{\pi_0^*} \right) \right\|_2^2 > \chi^2_{s,1-\alpha} \right\}$. *Then the procedure consisting of rejecting the null hypothesis* $\mathbb{H}_0$ *when*

$$\zeta_{N,T} := \frac{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Z^{(t)}) \mathbb{1}_{Z^{(t)} \in \widetilde{W}_N}}{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Z^{(t)})} > \alpha,$$

has an asymptotic level lower than α *in the sense that for any* $\epsilon > 0$, *there exists* $N_0 \in \mathbb{N}$ *such that for any* $N \geq N_0$ *it holds,*

$$\mathbb{P}\Big(\bigcup_{T_N \in \mathbb{N}} \bigcap_{T \geq T_N} \{\zeta_{N,T} \leq \alpha + \epsilon\} \Big) = 1.$$

6.6.2.2 Asymptotic confidence region

With Theorem 6.8, we proved that $\mathbf{X}_M^\top Y$ with Y distributed according to $\overline{\mathbb{P}}_{\pi^*}$ satisfies a CLT with an asymptotic Gaussian distribution centered at $\mathbf{X}_M^\top \overline{\pi}^{\pi^*}$. Using an approach analogous to Section 6.6.1.2, we propose here to build an asymptotic confidence region for π^*. The proof of Proposition 6.13 is postponed to Section 6.8.9.

Proposition 6.13. *We keep the notations and the assumptions of Theorem 6.8 and we consider $\alpha \in (0,1)$. We assume further that there exist $p \in [1,\infty]$ and $\kappa, R > 0$ such that*

$$\pi^* \in \mathbb{B}_p\big(\frac{\mathbf{1}_N}{2}, R\big) \quad \textit{and} \quad \forall \pi \in \mathbb{B}_p\big(\frac{\mathbf{1}_N}{2}, R\big), \quad \lambda_{\min}(\overline{\Gamma}^{\pi}) \geq \kappa.$$

Let us consider any estimator $\pi^\star \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R)$ of π^*. Then the probability of the event

$$\|\pi^* - \pi^\star\|_2 \leq (4\kappa)^{-1}\big\{\|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^\star})\|_2 + Cc^{-1}\sqrt{\chi^2_{s,1-\alpha}} + \|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star})\|_2\big\},$$

tends to $1-\alpha$ as $N \to \infty$.

Remarks.

- Analogously to Section 6.6.1.2, Proposition 6.13 motivates us to choose $\pi^\star$ among the minimizers of the function
$$M : \pi \mapsto \|\mathbf{X}_M^\top \overline{\pi}^{\pi} - \mathbf{X}_M^\top Y\|_2^2.$$
As mentioned in the Section 6.6.1.2, one can rely for example on a deep learning or a gradient descent method in order to reach a local minimum $\pi^\star$ for M.
- The term $\|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star})\|_2$ arising in the confidence region from Proposition 6.13 illustrates that our conditional CLT from Theorem 6.8 holds on $\mathbf{X}_M^\top Y$ and that we do not control what occurs in the orthogonal complement of the span of the columns of $\mathbf{X}_M$. Nevertheless, let us comment informally our result in the case where $E_M = \{0,1\}^N$ (meaning that there is no conditioning) and where ϑ^* is close to 0 (meaning that π^* is close to $\mathbf{1}_N/2$). In this framework, $\overline{\Gamma}^{\pi} = \mathrm{Diag}(\pi \odot (1-\pi))$ is close to $\frac{1}{4}\mathrm{Id}_N$ for π in a small neighbourhood around $\mathbf{1}_N/2$. Hence, we get that κ is approximately $\frac{1}{4}$. Since it also holds that $\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star} = \pi^* - \pi^\star$ (since $E_M = \{0,1\}^N$), we obtain from Proposition 6.13 that a CR for $\mathrm{Proj}_{\mathbf{X}_M}\pi^*$ with asymptotic coverage $1-\alpha$ is
$$\|\mathrm{Proj}_{\mathbf{X}_M}(\pi^* - \pi^\star)\|_2 \leq \|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^\star})\|_2 + Cc^{-1}\sqrt{\chi^2_{s,1-\alpha}}.$$

6.7 Numerical results

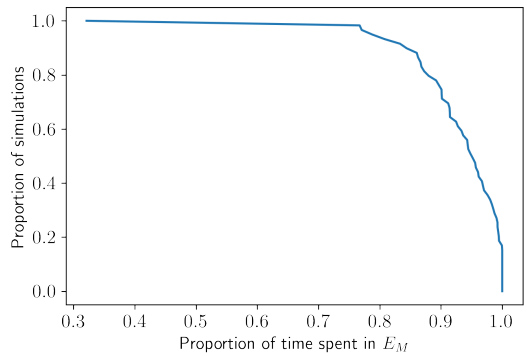
The code to reproduce our results is available at the following url: `https://github.com/quentin-duchemin/LogPSI`.

6.7.1 Sampling from the condition distribution

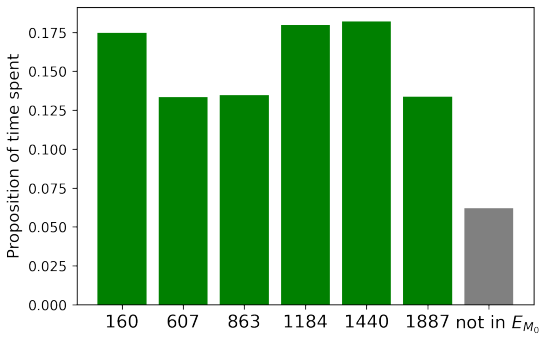
Description of the experiment. We test our approach under the global null, namely we consider $\vartheta^* = 0$. We work with $N = 11$, $p = 20$, $\lambda = 1.7$ and $\delta = 0.01$. The entries of the design matrix $\mathbf{X}$ are i.i.d. with a standard normal distribution. By choosing this toy example with a small value for N, we are able to compute exactly the selection event by running over the 2^N possible vectors $Y \in \{0,1\}^N$. We start by sampling some vector $Y_0 \in \{0,1\}^N$ with i.i.d. entries with a Bernoulli distribution of parameter $1/2$. Note the tuple $(\mathbf{X}, Y_0\lambda)$ determined the set M which corresponds to the set of indexes $i \in [d]$ such that $\hat{\vartheta}_i^\lambda \neq 0$ where $\hat{\vartheta}^\lambda$ is defined by Eq.(6.12). Given the equicorrelation set M, we run 40 simulated annealing paths of length $T = 150,000$ using the SEI-SLR algorithm (see Algorithm 5).

Uniform distribution on the selection event. In the following, we identify each vector $Y \in \{0,1\}^N$ with the number between 0 and $2^N - 1 = 2047$ that it represents in the base-2 numeral system. Using this identification, it holds on our example that $E_M = \{160, 607, 863, 1184, 1440, 1887\}$.

Figure 6.4.(a) shows the time spent in the selection event over the last $15,000$ time steps of the simulated annealing path for each of our 40 simulations. One can see that around 75% of the generated paths are spending at least 90% of their time in the selection event for the last $15,000$ time steps. Figure 6.4.(b) presents the proportion of time spent in the different states of the selection event and outside of E_M working again with the last $15,000$ visited states for each of the 40 simulations.

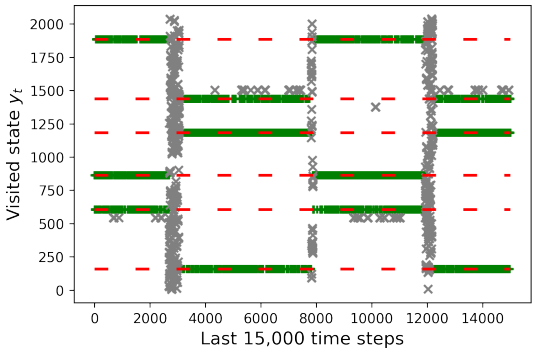


(a) Proportion of simulations (vertical axis) spending at least some $x\,\%$ of their time (horizontal axis) in the selection event.



(b) Time spent in each state of E_M and outside of E_M.

Figure 6.4: Visualization of the time spent in the selection event keeping the last $15,000$ visited states of each sequence provided by our algorithm SEI-SLR.



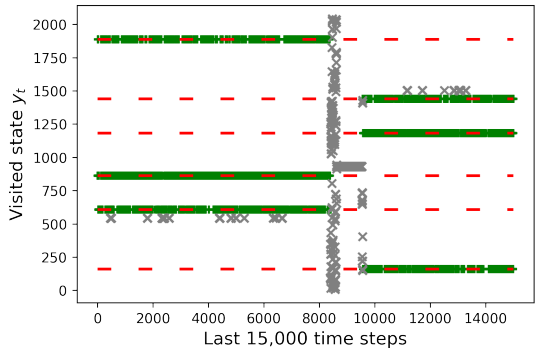


Figure 6.5: Last visited states of two simulated annealing paths. The dotted red lines indicate all the states belonging to the selection event. The gray (respectively green) crosses indicate visited states that do not belong (respectively that belong) to the selection event.

Figure 6.5 shows the last $15,000$ visited states for two specific simulated annealing paths. On the vertical axis, we have the integers encoded by all possible vectors $Y \in \{0,1\}^N$. The red dashed lines represent the states that belongs to the selection event E_M. While crosses are showing the visited states on the last $15,000$ time steps of the path, green crosses are emphasizing the ones that belong to the selection event. On this example, we see that the SEI-SLR algorithm covers properly the selection event without being stuck in one specific state of E_M. Each simulated annealing path is jumping from a state of E_M to another, ending up with an asymptotic distribution of the visited states that approximates the uniform distribution on E_M (see Figure 6.4.(b)). Let us point that two neighboring states in space $\{0,1\}^N$ will not necessarily be encoded by close integers.
Figure 6.5 suggests that the vectors encoded by the integers 160, 1184 and 1440 are close in the space $\{0,1\}^N$. Indeed, we see for example on the right plot of Figure 6.5 that in the last $5,000$ visited states, our algorithm goes from one of these three states to another passing through almost no state that does not belong to the selection event (this can be seen because there are only few gray crosses in the last $5,000$

iterations). The same remark holds for the three states encoded by the integers 607, 863 and 1887. However, we observe a large number of visited states that do not belong to E_M when we perform a transition from one of the state of the first group $\{160, 1184, 1440\}$ to one of the state of the second group $\{607, 863, 1887\}$. We can therefore legitimately think that the selection event separates into two groups of fairly distant states. This is confirmed by Figure 6.6 which presents the Hamming distances between the different vectors of E_M and reveals the existence of two clusters.

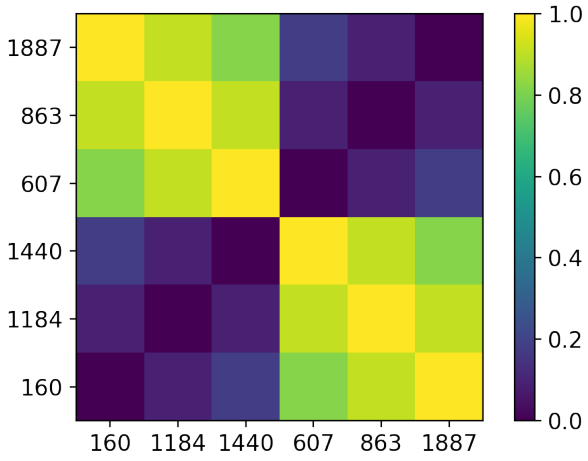


Figure 6.6: Normalized (by N) Hamming distances between the different states of the selection event. We observe two distinct clusters.

Comparison with the linear model. The previous theoretical and numerical results show that our approach allows to correctly identify the selection event E_M. Nevertheless, this method suffers from the curse of dimensionality since the random walks in the simulated annealings need to cover a state space of 2^N points. Let us mention that even in the linear model where the selection event E_M has the nice property to be a union of polyhedra, the method from Lee et al. [2016] to provide inference on a linear transformation of Y can also cope with some computational issues. Indeed, the construction of confidence intervals conditionally on the event E_M requires the computation of 2^s intervals (while the computation of each of them requires at least N^3 operations) where $s = |M|$ (see [Lee et al., 2016, Section 6]). Roughly speaking, both our approach in the logistic model and the one from [Lee et al., 2016, Section 6] in the linear model are limited in large dimensions. While in the linear case, computational efficiency of the known methods mainly depends on $s = |M|$, the extra cost arising from the non-linearity of the logistic model is their dependence on N.
Let us finally mention that in the Gaussian linear model (cf. Section 6.1.3), one can bypass the limitation of computing the 2^s intervals for each possible vector of dual signs on the equicorrelation set M by conditioning further on the observed vector of signs $\widehat{S}_M(Y) = \mathrm{sign}(\widehat{\theta}^\lambda)_M$. Stated otherwise, instead of conditioning on E_M, we condition on $E_M^{S_M}$ where $S_M = \widehat{S}_M(Y)$. This method reduces the computational burden but it will lead in general to less powerful inference procedures due to some information loss which can be quantified through the so-called left-over Fisher information. In Section 6.9, we discuss with further details PSI when we condition additionally on the observed vector of signs.

6.7.2 Hypothesis Testing

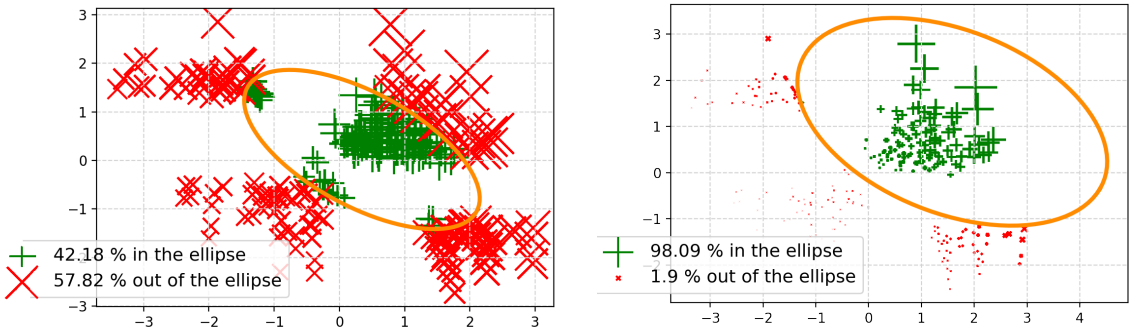
Description of the experiment. We consider $d = 40$, $N = 15$, $s = 2$, $\vartheta^* = [1, 1, 0, 0, \ldots, 0]$ and $\mathbf{X} \in \mathbb{R}^{N \times d}$ constructed as follows:

- We first sample $\widetilde{\mathbf{X}} \in \mathbb{R}^{N \times d}$ with independent standard normal entries.
- We normalize the first two columns of $\widetilde{\mathbf{X}}$ (i.e. we divide each entry of one column by the L^2-norm of this column) and the result gives the two first columns of $\mathbf{X}$.

- We project each column $j \in \{3, \dots, d\}$ onto $\mathrm{Span}(\mathbf{X}_1, \mathbf{X}_2)^\perp$. Thereafter, we normalize the resulting columns and we stack them to obtain the columns with index from 3 to d for the matrix $\mathbf{X}$.

Using this design matrix allows us to guarantee that the so-called mutual incoherence property [cf. Wainwright, 2019, Eq.(7.43b)] is satisfied. It is well known that this condition is used to ensure support recovery for the lasso [cf. Wainwright, 2019, Theorem 7.21] and by choosing a regularization parameter $\lambda = 1.4$, the selected support is equal to the true support $\widehat{M} = \{1, 2\}$. Working with $s = 2$, we can easily visualize the results of our multiple testing procedure. For any $\nu > 0$, we consider $\theta_0^* = [\nu, \nu]$ and we consider the hypothesis test (6.33).

Results. Figure 6.7 shows the way we compute the test statistic from Proposition 6.12: each visited state $Z^{(t)}$ of the SEI-SLR algorithm is plotted using a different color depending on whether $Z^{(t)} \in \widetilde{W}_N$ or $Z^{(t)} \notin \widetilde{W}_N$. Each sample is weighted proportionally to $\mathbb{P}_{\theta_0^*}(Z^{(t)})$ so that we reject the null if and only if the sum of weights of samples falling outside of the orange ellipse is larger than ν times the total mass of samples. In Figure 6.8, we show for ν ranging from 0 to 2 the total mass of samples falling into the ellipse $\left(\widetilde{W}_N\right)^c$. We see that our test controls that type I error at level $\alpha = 5\%$. Moreover, the test seems much more powerful when $\nu < 1$ compared to the case where $\nu > 1$.



(a) $\theta_0^* = [0.1, 0.1]$ (b) $\theta_0^* = [0.9, 0.9]$

Figure 6.7: The orange ellipse represents the set of parameter $\theta \in \mathbb{R}^s$ such that $\|\widetilde{G}_N^{-1/2} G_N(\widetilde{\theta})(\theta - \widetilde{\theta})\|_2^2 = \chi^2_{s,1-\alpha}$. For each t, we plot the MLE $\Psi(\mathbf{X}_M^\top Y^{(t)})$ with a green plus if the point falls into the orange ellipse and with a red cross otherwise. The size of the markers is proportional to $\overline{\mathbb{P}}_{\theta_0^*}(Y^{(t)})$.

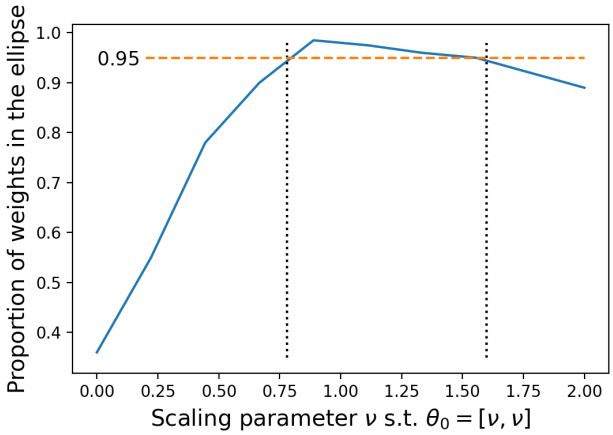


Figure 6.8: Value of the test statistic for $\theta_0^* = [\nu, \nu]$ with ν ranging from 0 to 2. The dashed vertical lines show the values of ν so that we reject the null at level 5%.

Let us highlight that we trained a feed-forward neural network with three hidden layers to approximate $\Psi = \Xi^{-1}$ where we recall the $\Xi(\theta) = \mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta)$ for any $\theta \in \mathbb{R}^s$.

6.7.3 Confidence region

As presented in Proposition 6.11 and the subsequent remark, the size of our confidence region is mainly driven by the distance $\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2$. This encourages us to choose our estimate $\widehat{\theta}^*$ among the local minimizers of the function $m : \theta \mapsto \|\overline{\theta}(\theta) - \widehat{\theta}\|_2$. In the following, we propose a deep-learning and a gradient descent approach to achieve this goal.

6.7.3.1 Deep learning method

We train a feed forward neural network with ReLu activation function and three hidden layers. With this network, we aim at estimating any $\theta \in \mathbb{R}^s$ by feeding as input $\overline{\theta}(\theta)$. We generate our training dataset by first sampling $n_{train} = 500$ random vectors $\theta_i \sim \mathcal{N}(0, \mathrm{Id}_s)$, $i \in [n_{train}]$. Then, for any $i \in [n_{train}]$ we compute the estimate $\widetilde{\theta}(\theta_i)$ of $\overline{\theta}(\theta_i)$ as follows

$$\widetilde{\pi}^{\theta_i} = \frac{\sum_{t=1}^T \mathbb{P}_{\theta_i}(Y^{(t)}) Y^{(t)}}{\sum_{t=1}^T \mathbb{P}_{\theta_i}(Y^{(t)})} \quad \text{and} \quad \widetilde{\theta}(\theta_i) = \Psi(\mathbf{X}_M^\top \widetilde{\pi}^{\theta_i}),$$

where $(Y^{(t)})_{t \geq 1}$ is the sequence generated from the SEI-SLR algorithm (see Algorithm 5). We train our network using stochastic gradient descent with learning rate 0.01 and 500 epochs. At each epoch, we feed to the network the inputs $(\widetilde{\theta}(\theta_i))_{i \in [n_{train}]}$ with the corresponding target values $(\theta_i)_{i \in [n_{train}]}$. We then compute our estimate $\theta^\bigstar$ of θ^* by taking the output of our network when taking as input the unpenalized MLE $\widehat{\theta}$ using the design $\mathbf{X}_M$ (cf. Eq.(6.28)). Figure 6.9 illustrates the result obtained from this deep learning approach. We keep the experiment settings of Section 6.7.2 namely, we consider $\vartheta^* = (1\ 1\ 0\ \ldots\ 0)^\top \in \mathbb{R}^d$ and we choose the regularization parameter λ so that the selected model corresponds to the true set of active variables, namely $M = \{1, 2\}$.

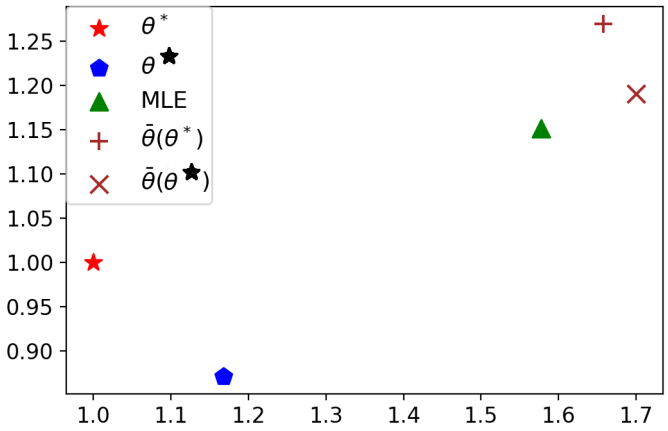


Figure 6.9: Visualization of the results obtained using our deep learning approach to compute an estimate $\theta^\bigstar$ (the blue hexagone) of θ^* (the red star). $\theta^\bigstar$ corresponds to the output of the neural network when feeding as input the MLE $\widehat{\theta}$ (the green triangle). We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

6.7.3.2 Gradient descent method

As shown in the proof of the expression of Proposition 6.11 (cf. Eq.(6.53)), it holds

$$\forall \theta \in \mathbb{R}^s, \quad \nabla_\theta \overline{\pi}^\theta = \overline{\Gamma}^\theta \mathbf{X}_M.$$

Recalling additionally that $\overline{\theta}(\theta) = \Psi\left(\mathbf{X}_M^\top \overline{\pi}^\theta\right)$ (cf. Eq.(6.30)), we get that for any $\theta \in \mathbb{R}^s$,

$$\begin{aligned}\nabla_\theta m(\theta) &= 2\nabla_\theta \overline{\theta}(\theta)(\overline{\theta}(\theta) - \widehat{\theta}) \\ &= 2\nabla\Psi(\mathbf{X}_M^\top \overline{\pi}^\theta)\mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}) \\ &= 2\nabla\Psi(\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta)})\mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}) \\ &= 2\left(\mathbf{X}_M^\top \mathrm{Diag}(\pi^{\overline{\theta}(\theta)} \odot (1 - \pi^{\overline{\theta}(\theta)}))\mathbf{X}_M\right)^{-1} \mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}).\end{aligned}$$

Hence,

$$\nabla_\theta m(\theta) = 2\left[G_N(\overline{\theta}(\theta))\right]^{-1} \mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}).$$

Given some θ, $\overline{\pi}^\theta$ and $\overline{\Gamma}^\theta$ can be estimated using samples generated by the SEI-SLR algorithm (and thus the same holds for $\overline{\theta}(\theta) = \Psi(\mathbf{X}_M^\top \overline{\pi}^\theta)$ and for $G_N(\overline{\theta}(\theta))$).

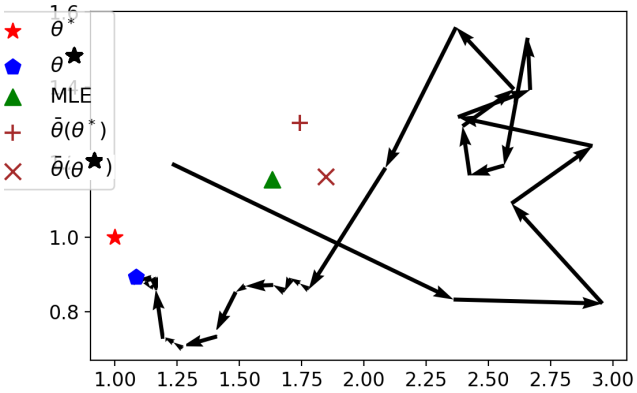


Figure 6.10: Visualization of our gradient descent procedure to compute an estimate $\theta^\bigstar$ (the blue hexagone) of θ^* (the red star). The MLE $\widehat{\theta}$ is the green triangle. We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

6.8 Proofs

6.8.1 Proof of Proposition 6.1

Let us consider ϑ_1, ϑ_2 two vectors in $\mathbb{R}^d$ achieving the minimum in (6.12). Then, denoting $\vartheta_3 = \frac{1}{2}\vartheta_1 + \frac{1}{2}\vartheta_2$ it holds

$$\frac{\mathcal{L}_N(\vartheta_1, Z) + \mathcal{L}_N(\vartheta_2, Z)}{2} + \lambda\frac{\|\vartheta_1\|_1 + \|\vartheta_2\|_1}{2} \leq \mathcal{L}_N(\vartheta_3, Z) + \lambda\|\vartheta_3\|_1.$$

Since the triangle inequality gives $\|\vartheta_3\|_1 \leq \frac{\|\vartheta_1\|_1 + \|\vartheta_2\|_1}{2}$ and since the function ξ is strictly convex, it holds that $\mathbf{X}\vartheta_1 = \mathbf{X}\vartheta_2$. Indeed, otherwise we would have by strict convexity

$$\begin{aligned}& \mathcal{L}_N(\vartheta_3, Z) + \lambda\|\vartheta_3\|_1 \\ = & \sum_{i=1}^N (\xi(\langle \mathbf{x}_i, \vartheta_3\rangle) - \langle y_i\mathbf{x}_i, \vartheta_3\rangle) + \lambda\|\vartheta_3\|_1 \\ \leq & \sum_{i=1}^N \left(\xi(\langle \mathbf{x}_i, \frac{\vartheta_1 + \vartheta_2}{2}\rangle) - \frac{1}{2}\langle y_i\mathbf{x}_i, \vartheta_1\rangle - \frac{1}{2}\langle y_i\mathbf{x}_i, \vartheta_2\rangle\right) + \frac{1}{2}\lambda\|\vartheta_1\|_1 + \frac{1}{2}\lambda\|\vartheta_2\|_1 \\ < & \frac{\mathcal{L}_N(\vartheta_1, Z) + \mathcal{L}_N(\vartheta_2, Z)}{2} + \lambda\frac{\|\vartheta_1\|_1 + \|\vartheta_2\|_1}{2}.\end{aligned}$$

From the KKT conditions, we deduce that for a given $Y \in \mathcal{Y}^N$, all solutions $\hat{\vartheta}^\lambda$ of (6.12) have the same vector of signs denoted $\widehat{S}(Y)$ which is given

$$\widehat{S}(Y) = \frac{1}{\lambda}\mathbf{X}^\top \left(Y - \sigma(\mathbf{X}\hat{\vartheta}^\lambda)\right),$$

where $\hat{\vartheta}^\lambda$ is any solution to (6.12).

6.8.2 Proof of Proposition 6.2

Partitioning the KKT conditions of Eq.(6.13) according to the equicorrelation set $\widehat{M}(Y)$ leads to

$$\begin{aligned}
\mathbf{X}_{\widehat{M}(Y)}^\top \left(Y - \sigma(\mathbf{X}_{\widehat{M}(Y)}\hat{\vartheta}^\lambda_{\widehat{M}(Y)})\right) &= \lambda \widehat{S}_{\widehat{M}(Y)} \\
\mathbf{X}_{-\widehat{M}(Y)}^\top \left(Y - \sigma(\mathbf{X}_{\widehat{M}(Y)}\hat{\vartheta}^\lambda_{\widehat{M}(Y)})\right) &= \lambda \widehat{S}_{-\widehat{M}(Y)} \\
\operatorname{sign}(\hat{\vartheta}^\lambda_{\widehat{M}(Y)}) &= \widehat{S}_{\widehat{M}(Y)} \\
\|\widehat{S}_{-\widehat{M}(Y)}\|_\infty &< 1
\end{aligned}$$

Since the KKT conditions are necessary and sufficient for a solution, we obtain that Y belongs to $E_M^{S_M}$ if and only if there exists $\theta \in \mathbb{R}^s$ satisfying

$$\begin{aligned}
\mathbf{X}_M^\top \left(Y - \sigma(\mathbf{X}_M\theta)\right) &= \lambda S_M \\
\operatorname{sign}(\theta) &= S_M \\
\|\mathbf{X}_{-M}^\top \left(Y - \sigma(\mathbf{X}_M\theta)\right)\|_\infty &< \lambda
\end{aligned}$$

6.8.3 Proof of Proposition 6.3

Let us consider $\theta, \theta' \in \mathbb{R}^s$ such that $\Xi(\theta) = \Xi(\theta')$. Then we have

$$\begin{aligned}
0 &= \mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta) - \mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta') \\
&= \Xi(\theta) - \Xi(\theta') \\
&= \int_0^1 \nabla\Xi(\theta t + (1-t)\theta') \cdot (\theta - \theta')dt \\
&= \int_0^1 \mathbf{X}_M^\top \operatorname{Diag}\left[\sigma'(\mathbf{X}_M\theta t + (1-t)\mathbf{X}_M\theta')\right]\mathbf{X}_M(\theta - \theta')dt \\
&= \mathbf{X}_M^\top \underbrace{\left(\int_0^1 \operatorname{Diag}\left[\sigma'(\mathbf{X}_M\theta t + (1-t)\mathbf{X}_M\theta')\right]dt\right)}_{=:D} \mathbf{X}_M(\theta - \theta'). \qquad (6.35)
\end{aligned}$$

Note that for any $t \in [0,1]$ and for any $i \in [N]$, $\{\sigma'(\mathbf{X}_M\theta t + (1-t)\mathbf{X}_M\theta')\}_i > 0$ since $\xi''(u) = \sigma'(u) > 0$ for any $u \in \mathbb{R}$. We deduce that $D \in \mathbb{R}^{N\times N}$ is a diagonal matrix with strictly positive coefficients on the diagonal. Eq.(6.35) gives that $\theta - \theta' \in \operatorname{Ker}(\mathbf{X}_M^\top D\mathbf{X}_M)$ which implies that $(\theta-\theta')^\top \mathbf{X}_M^\top D\mathbf{X}_M(\theta-\theta') = 0$. This means that

$$\sum_{i=1}^N D_{i,i}\left[\mathbf{X}_M(\theta - \theta')\right]_i^2 = 0.$$

Since $D_{i,i} > 0$ for all $i \in [N]$, we get that $\mathbf{X}_M(\theta - \theta') = 0$, i.e. $\mathbf{X}_M\theta = \mathbf{X}_M\theta'$. Since $\mathbf{X}_M$ has full column rank, this leads to $\theta = \theta'$.
Since Ξ is injective and of class $\mathcal{C}^m$ with a differential given by $\nabla_\theta\Xi(\theta) = \mathbf{X}_M^\top \operatorname{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M$ which is invertible at any $\theta \in \mathbb{R}^s$ under the assumptions of Proposition 6.3 Hence the global inversion theorem gives Proposition 6.3.

6.8.4 Proof of Theorem 6.8

For the sake of brevity, we will simply denote $\overline{G}_N(\pi^*)$ by $\overline{G}_N$. Let us further denote $\mathbf{X}_M^\top = [\mathbf{w}_1 \mid \mathbf{w}_2 \mid \ldots \mid \mathbf{w}_N]$, where $\mathbf{w}_i = \mathbf{x}_{i,M} \in \mathbb{R}^s$. The proof of Theorem 6.8 relies on [Bardet et al., 2008, Theorem 1]. In the following, we check that all the assumptions of [Bardet et al., 2008, Theorem 1] are satisfied. Denoting for any $i \in [N]$, $\xi_{i,N} = \overline{G}_N^{-1/2}\mathbf{w}_i(y_i - \overline{\pi}_i^{\pi^*})$, it holds

$$\overline{G}_N^{-1/2}\mathbf{X}_M^\top(Y - \overline{\pi}^{\pi^*}) = \sum_{i=1}^N \overline{G}_N^{-1/2}\mathbf{w}_i(y_i - \overline{\pi}_i^{\pi^*}) = \sum_{i=1}^N \xi_{i,N}.$$

Let us also point that $\overline{\mathbb{E}}_{\pi^*}[\xi_{i,N}] = 0$. In the following, we will simply refer to $\xi_{i,N}$ as ξ_i to ease the reading of the proof. Let us denote further

$$A_N = \sum_{i=1}^N \overline{\mathbb{E}}_{\pi^*}\left(\|\xi_i\|_2^3\right).$$

One can notice that

$$\begin{aligned}\overline{\mathbb{E}}_{\pi^*}\left(\|\xi_i\|_2^3\right) &= \overline{\mathbb{E}}_{\pi^*}\left[(y_i - \overline{\pi}_i^{\pi^*})^3\right]\|\overline{G}_N^{-1/2}\mathbf{w}_i\|_2^3 \\ &\leq \left(\frac{K}{\sqrt{c}\overline{\sigma}_{\min}}\right)^3 N^{-3/2}s^{3/2},\end{aligned}$$

where we used that

$$\|\overline{G}_N^{-1/2}\mathbf{w}_i\|_2^2 \leq \|\overline{G}_N^{-1/2}\|^2 \times \|\mathbf{w}_i\|_2^2 \leq \|\overline{G}_N^{-1}\|(sK^2) \leq (c\overline{\sigma}_{\min}^2 N)^{-1}(sK^2).$$

We deduce that

$$A_N \leq \left(\frac{K}{\sqrt{c}\overline{\sigma}_{\min}}\right)^3 N^{-1/2}s^{3/2}.$$

Hence $A_N \underset{N\to\infty}{\to} 0$ which the first condition that needed to be checked to apply [Bardet et al., 2008, Theorem 1].
Let us now check the second condition from that Bardet et al. [2008] that consists in identifying the appropriate asymptotic covariance matrix.

$$\begin{aligned}\sum_{i=1}^N \overline{Cov}_{\pi^*}(\xi_i) &= \sum_{i=1}^N \overline{\mathbb{E}}_{\pi^*}\left[\overline{G}_N^{-1/2}\mathbf{w}_i\mathbf{w}_i^\top\overline{G}_N^{-1/2}(y_i - \overline{\pi}_i^{\pi^*})^2\right] \\ &= \sum_{i=1}^N \overline{G}_N^{-1/2}\mathbf{w}_i \underbrace{\overline{\mathbb{E}}_{\pi^*}(y_i - \overline{\pi}_i^{\pi^*})^2}_{=(\overline{\sigma}_i^{\pi^*})^2}\mathbf{w}_i^\top\overline{G}_N^{-1/2} \\ &= \overline{G}_N^{-1/2}\sum_{i=1}^N \mathbf{w}_i(\overline{\sigma}_i^{\pi^*})^2\mathbf{w}_i^\top\overline{G}_N^{-1/2} \\ &= \overline{G}_N^{-1/2}\mathbf{X}_M^\top\mathrm{Diag}\big((\overline{\sigma}^{\pi^*})^2\big)\mathbf{X}_M\overline{G}_N^{-1/2} \\ &= \overline{G}_N^{-1/2}\overline{G}_N\overline{G}_N^{-1/2} \\ &= \mathrm{Id}_s.\end{aligned}$$

To apply [Bardet et al., 2008, Theorem 1], it remains to check that the dependent Lindeberg conditions hold. For this, we consider some map $f \in \mathcal{C}_b^3(\mathbb{R}^s, \mathbb{R})$ where $\mathcal{C}_b^3(\mathbb{R}^s, \mathbb{R})$ is the set of functions from $\mathbb{R}^s$ to $\mathbb{R}$ with bounded and continuous partial derivatives up to order 3. In the following, we denote

$$W_i = \overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top(Y - \overline{\pi}^{\pi^*})_{[i-1]} = \sum_{a=1}^{i-1}\xi_a.$$

First dependent Lindeberg condition.

For any $i \in [N]$, let us consider W_i' (resp. ξ_i') an independent copy of the random vector W_i (resp. ξ_i). Let us recall the following well-known result

Lemma 6.14. *Let us consider two real valued random variables A, B on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let us consider (A', B') an independent copy of the random vector (A, B). Then it holds,*

$$Cov(A, B) = \frac{1}{2}\mathbb{E}\big[(A - A')(B - B')\big].$$

Using Lemma 6.14, the Cauchy-Schwarz inequality and Jensen's inequalities, we get,

$$\begin{aligned}
&\sum_{k,l=1}^{s}\sum_{i=1}^{N}|\overline{Cov}_{\pi^*}(\frac{\partial^2 f}{\partial x_l \partial x_k}(W_i), (\xi_i)_k(\xi_i)_l)| \\
&= \sum_{k,l=1}^{s}\sum_{i=1}^{N}|\overline{Cov}_{\pi^*}(\frac{\partial^2 f}{\partial x_l \partial x_k}(W_i), (\xi_i)_k(\xi_i)_l)| \\
&= \sum_{k,l=1}^{s}\sum_{i=1}^{N}\frac{1}{2}|\overline{\mathbb{E}}_{\pi^*}\left[\left(\frac{\partial^2 f}{\partial x_l \partial x_k}(W_i) - \frac{\partial^2 f}{\partial x_l \partial x_k}(W_i')\right)((\xi_i)_k(\xi_i)_l - (\xi_i')_k(\xi_i')_l)\right]| \\
&\leq \sum_{k,l=1}^{s}\sum_{i=1}^{N}\frac{1}{2}\|\nabla^3 f\|_\infty \overline{\mathbb{E}}_{\pi^*}(\|W_i - W_i'\|_2 \times |(\xi_i)_k(\xi_i)_l - (\xi_i')_k(\xi_i')_l|) \\
&\leq \sum_{k,l=1}^{s}\sum_{i=1}^{N}\frac{1}{2}\|\nabla^3 f\|_\infty \sqrt{\overline{\mathbb{E}}_{\pi^*}(\|W_i - W_i'\|_2^2)} \times \sqrt{\overline{\mathbb{E}}_{\pi^*}(|(\xi_i)_k(\xi_i)_l - (\xi_i')_k(\xi_i')_l|^2)} \\
&\leq \sum_{k,l=1}^{s}\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\overline{\mathbb{V}ar}_{\pi^*}(\|W_i\|_2)} \times \sqrt{\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_k(\xi_i)_l|)} \\
&\leq s\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\overline{\mathbb{V}ar}_{\pi^*}(\|W_i\|_2)} \times \sqrt{\sum_{k,l=1}^{s}\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_k(\xi_i)_l|)},
\end{aligned}$$

where in the last inequality we used Jensen's inequality. Let us upper-bound the terms $\overline{\mathbb{V}ar}_{\pi^*}(\|W_i\|_2)$ and $\sum_{k,l=1}^{s}\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_k(\xi_i)_l|)$ independently. We have

$$\begin{aligned}
&\overline{\mathbb{V}ar}_{\pi^*}(\|W_i\|_2) \\
&\leq \overline{\mathbb{E}}_{\pi^*}\left(\|W_i\|_2^2\right) \\
&= \overline{\mathbb{E}}_{\pi^*}\left[(Y - \overline{\pi}^{\pi^*})_{[i-1]}^\top \mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top (Y - \overline{\pi}^{\pi^*})_{[i-1]}\right] \\
&= \overline{\mathbb{E}}_{\pi^*}\left[\mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top (Y - \overline{\pi}^{\pi^*})_{[i-1]}(Y - \overline{\pi}^{\pi^*})_{[i-1]}^\top \mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right)\right] \\
&= \mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{k,l=1}^{s}\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_k(\xi_i)_l|) \\
&= \sum_{k,l=1}^{s}((\overline{G}_N^{-1/2})_{k,:}\mathbf{w}_i)^2((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2\left\{\overline{\mathbb{E}}_{\pi^*}\left[(y_i - \overline{\pi}_i^{\pi^*})^4\right] - \overline{\mathbb{E}}_{\pi^*}\left[(y_i - \overline{\pi}_i^{\pi^*})^2\right]^2\right\} \\
&= \sum_{k,l=1}^{s}((\overline{G}_N^{-1/2})_{k,:}\mathbf{w}_i)^2((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2(\overline{\sigma}_i^{\pi^*})^2(1 - 2\overline{\pi}_i^{\pi^*})^2 \\
&= \|\overline{G}_N^{-1/2}\mathbf{w}_i\|_2^4(\overline{\sigma}_i^{\pi^*})^2(1 - 2\overline{\pi}_i^{\pi^*})^2 \\
&\leq K^4(c\overline{\sigma}_{\min}^2)^{-2}\frac{s^2}{N^2}(\overline{\sigma}_i^{\pi^*})^2(1 - 2\overline{\pi}_i^{\pi^*})^2,
\end{aligned}$$

where $(\overline{\sigma}_i^{\pi^*})^2 = \overline{\pi}_i^{\pi^*}(1-\overline{\pi}_i^{\pi^*})$. Hence, coming back the first Lindeberg condition, we have (forgetting to mention the constants $K, s, c, \overline{\sigma}^2_{\min}$ that do not depend on N, which is the sense of the symbol $\lesssim$),

$$
\begin{aligned}
&\sum_{k,l=1}^{s}\sum_{i=1}^{N}|\overline{Cov}_{\pi^*}(\frac{\partial^2 f}{\partial x_l \partial x_k}(W_i),(\xi_i)_k(\xi_i)_l)| \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right)(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\|\overline{G}_N^{-1}\|_F\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\frac{1}{N}\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\theta^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\leq \frac{1}{N^{3/2}}\|\nabla^3 f\|_\infty \sum_{i=1}^{N}\sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2},
\end{aligned}
$$

where we used that $\|G_N^{-1}\|_F \leq \sqrt{s}\|G_N^{-1}\| \lesssim N^{-1}$ (since $\overline{G}_N^{-1}$ has rank s, see Section 6.5.1). Hence, the first dependent Lindeberg condition from Bardet et al. [2008] holds thanks to the assumptions made in Theorem 6.8.

Second dependent Lindeberg condition.
Using an approach analogous to the one conducted for the first dependent Lindeberg condition, one can obtain

$$
\begin{aligned}
&\sum_{l=1}^{s}\sum_{i=1}^{N}|\overline{Cov}_{\pi^*}(\frac{\partial f}{\partial x_l}(W_i),(\xi_i)_l)| \\
&\leq \sqrt{s}\sum_{i=1}^{N}\|\nabla^2 f\|_\infty \sqrt{\overline{\mathbb{V}ar}_{\pi^*}(\|W_i\|_2)} \times \sqrt{\sum_{l=1}^{s}\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_l|)} \\
&\lesssim \frac{1}{\sqrt{N}}\|\nabla^2 f\|_\infty \sum_{i=1}^{N}\sqrt{\mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right)\left(1-2\overline{\pi}_i^{\pi^*}\right)^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\lesssim \frac{1}{\sqrt{N}}\|\nabla^2 f\|_\infty \sum_{i=1}^{N}\sqrt{\|\overline{G}_N^{-1}\|_F\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F\left(1-2\overline{\pi}_i^{\pi^*}\right)^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\lesssim \frac{1}{N}\|\nabla^2 f\|_\infty \sum_{i=1}^{N}\sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F\left(1-2\overline{\pi}_i^{\pi^*}\right)^2(\overline{\sigma}_i^{\pi^*})^2},
\end{aligned}
$$

where we used that

$$
\begin{aligned}
&\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_l|) \\
&= \overline{\mathbb{E}}_{\pi^*}\left(|(\xi_i)_l|^2\right) - \left(\overline{\mathbb{E}}_{\pi^*}|(\xi_i)_l|\right)^2 \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2\left\{\overline{\mathbb{E}}_{\pi^*}\left((y_i-\overline{\pi}_i^{\pi^*})^2\right) - \left(\overline{\mathbb{E}}_{\pi^*}|y_i-\overline{\pi}_i^{\pi^*}|\right)^2\right\} \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2\left\{\overline{\pi}_i^{\pi^*}(1-\overline{\pi}_i^{\pi^*}) - \left(\overline{\pi}_i^{\pi^*}(1-\overline{\pi}_i^{\pi^*}) + (1-\overline{\pi}_i^{\pi^*})\overline{\pi}_i^{\pi^*}\right)^2\right\} \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2\overline{\pi}_i^{\pi^*}(1-\overline{\pi}_i^{\pi^*})\left(1-4(1-\overline{\pi}_i^{\pi^*})\overline{\pi}_i^{\pi^*}\right) \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2(\overline{\sigma}_i^{\pi^*})^2\left(1-2\overline{\pi}_i^{\pi^*}\right)^2 \\
&\lesssim \frac{1}{N}(\overline{\sigma}_i^{\pi^*})^2\left(1-2\overline{\pi}_i^{\pi^*}\right)^2.
\end{aligned}
$$

Assuming that

$$\sum_{i=1}^{N} \sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]} \mathbf{X}_{[i-1],M}\|_F \left(1 - 2\overline{\pi}_i^{\pi^*}\right)^2} \underset{N\to\infty}{=} o(N),$$

we obtain applying [Bardet et al., 2008, Theorem 1] the following CLT

$$\overline{G}_N^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*}) \underset{N\to+\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}(0, \mathrm{Id}_s).$$

6.8.5 Proof of Theorem 6.9

To make the notations less cluttered, we will simply denote in the following $\overline{G}_N(\theta^*)$ by $\overline{G}_N$ and $\overline{\theta}(\theta^*)$ by $\overline{\theta}$.

First step. We use Theorem 6.8 where we established a CLT for

$$-L_N(\overline{\theta}, (Y, \mathbf{X}_M)) = \mathbf{X}_M^\top (Y - \pi^{\overline{\theta}}) = \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*}) = \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*}).$$

Let us highlight that the first equality comes directly from the definition of $L_N(\overline{\theta}, (Y, \mathbf{X}_M))$ (see Section 6.3.2), the second equality comes from Eq.(6.30) and the last equality holds since we work under the selected model meaning that $\pi^* = \sigma(\mathbf{X}\vartheta^*) = \sigma(\mathbf{X}_M\theta^*)$ (and thus that $\overline{\mathbb{P}}_{\theta^*} \equiv \overline{\mathbb{P}}_{\pi^*}$). Let us recall that to prove Theorem 6.8, we used a variant of the Linderberg CLT for dependent random variables proved by Bardet et al. [2008]. The proof of Theorem 6.8 is given in Section 6.8.4.

Second step. We now prove that for any $\epsilon > 0$ there is some $\delta > 0$ such that when N is large enough

$$\overline{\mathbb{P}}_{\theta^*}\left(\text{there is } \widehat{\theta} \in \mathcal{N}_N(\overline{\theta}, \delta) \text{ such that } L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0\right) > 1 - \epsilon,$$

with $\mathcal{N}_N(\overline{\theta}, \delta) = \{\theta \ : \ \|\overline{G}_N^{1/2}(\theta - \overline{\theta})\|_2 \leq \delta\}$. Stated otherwise, we will prove that there exist a constant $\delta > 0$ and an integer $N_\delta \in \mathbb{N}$ such that for any $N \geq N_\delta$, the following holds with high probability,

- the conditional MLE $\widehat{\theta}$ exists,
- the conditional MLE $\widehat{\theta}$ is contained in the ellipsoid $\mathcal{N}_N(\overline{\theta}, \delta)$ centered at $\overline{\theta}$.

Let us denote

$$\begin{aligned} F : \theta \in \mathbb{R}^s \mapsto & \overline{G}_N^{-1/2}(L_N(\overline{\theta}, (Y, \mathbf{X}_M)) - L_N(\theta, (Y, \mathbf{X}_M))) \\ & = \overline{G}_N^{-1/2} \mathbf{X}_M^\top (\pi^{\overline{\theta}} - \pi^\theta). \end{aligned}$$

Note that F is a deterministic function and does not depend on the random variable Y. Moreover we choose to leave implicit the dependence on N of F. We also point out that it holds for any $\theta \in \mathbb{R}^s$,

$$\nabla_\theta F(\theta) = -\overline{G}_N^{-1/2} \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta)) \mathbf{X}_M = -\overline{G}_N^{-1/2} G_N(\theta).$$

Hence F is a $\mathcal{C}^1$ map with invertible Jacobian at any $\theta \in \mathbb{R}^s$ and is injective (thanks to Proposition 6.3). Applying the global inversion theorem, we deduce that F is a $\mathcal{C}^1$-diffeomorphism from $\mathbb{R}^s$ to $\mathbb{R}^s$.

Sketch of proof.
In the following, we prove that for any ϵ, we can choose $\delta > 0$ such that for some $N_\delta \in \mathbb{N}$ and for any $N \geq N_\delta$, it holds on some event E_N satisfying $\overline{\mathbb{P}}_{\theta^*}(E_N) \geq 1 - \epsilon$,

$$\begin{aligned} & \overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M)) \in F(\mathcal{N}_N(\overline{\theta}, \delta)) \\ \Leftrightarrow \quad & \overline{G}_N^{-1/2} (\underbrace{\mathbf{X}_M^\top \overline{\pi}^{\theta^*}}_{=\mathbf{X}_M^\top \pi^{\overline{\theta}}} - \mathbf{X}_M^\top Y) \in F(\mathcal{N}_N(\overline{\theta}, \delta)). \end{aligned} \tag{6.36}$$

This would mean (by definition of F) that on E_N, there exists some $\widehat{\theta} \in \mathcal{N}_N(\overline{\theta}, \delta)$ such that $\overline{G}_N^{-1/2} L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0$ or equivalently that $L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0$. A sufficient condition for Eq.(6.36) to hold is to check that on the event E_N it holds

$$\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 < \inf_{\theta \in \partial \mathcal{N}_N(\overline{\theta}, \delta)} \|F(\theta)\|_2, \tag{6.37}$$

where $\partial \mathcal{N}_N(\overline{\theta}, \delta) := \{\theta \in \mathbb{R}^s \,|\, \|\overline{G}_N^{1/2}(\theta - \overline{\theta})\|_2 = \delta\}$. This sufficient condition is a direct consequence of Lemma 6.15 and Figure 6.11 gives a visualization of our proof strategy.

Lemma 6.15. *Let $f : \mathbb{R}^s \to \mathbb{R}^s$ be a $\mathcal{C}^1$-diffeomorphism from $\mathbb{R}^s$ to $f(\mathbb{R}^s)$. Then for any closed space $D \subset \mathbb{R}^s$ it holds*

$$f(\partial D) = \partial f(D),$$

where for any set $U \subseteq \mathbb{R}^s$, $\partial U = \overline{U} \backslash \mathring{U}$ with $\overline{U}$ the closure of the set U and $\mathring{U}$ the interior of the set U.

Proof. As a $\mathcal{C}^1$-diffeomorphism, f is in particular a homeomorphism, and as such, it preserves the topological structures. □

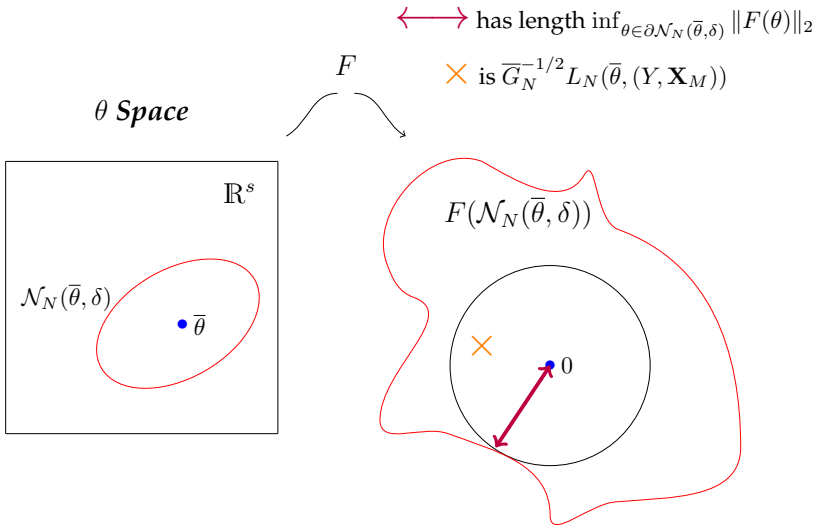


Figure 6.11: Visualization support for the proof of the existence of the MLE with large probability in a neighbourhood of $\overline{\theta}$. We show that with large probability, the orange cross is in the black circle (*i.e.*, Eq.(6.37) holds) which implies that the orange cross belongs to $F(\mathcal{N}_N(\overline{\theta}, \delta))$ (*i.e.*, Eq.(6.36) holds). The MLE is then defined as $\widehat{\theta} = F^{-1}(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M)) \in \mathcal{N}_N(\overline{\theta}, \delta)$.

Let $\epsilon > 0$ and let us consider

$$\delta := \frac{\mathfrak{K}^{1/2}}{\epsilon^{1/2} 2C^{-1} c \overline{\sigma}^2_{\min}}, \tag{6.38}$$

(the reason of this choice will become clear with Eq.(6.43)). Let us first notice that for any $\theta \in \mathbb{R}^s$,

$$\begin{aligned} & L_N(\overline{\theta}, (Y, \mathbf{X}_M)) - L_N(\theta, (Y, \mathbf{X}_M)) && (6.39)\\ & = \mathbf{X}_M^\top (\pi^{\overline{\theta}} - \pi^\theta) && (6.40)\\ & = \underbrace{\int_0^1 G_N(t\overline{\theta} + (1-t)\theta) dt}_{=: Q_N(\theta)} (\overline{\theta} - \theta), && (6.41) \end{aligned}$$

where we used that the Jacobian of the map $\theta \mapsto \mathbf{X}_M^\top \pi^\theta = \mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta)$ is $\mathbf{X}_M \mathrm{Diag}(\sigma'(\mathbf{X}_M \theta)) \mathbf{X}_M =$

$G_N(\theta)$. Recalling further that $\|\overline{G}_N^{-1/2}(\theta-\overline{\theta})\|_2 = \delta$ for any $\theta \in \partial\mathcal{N}_N(\overline{\theta},\delta)$, it holds,

$$
\begin{aligned}
&\inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \|F(\theta)\|_2 \\
&= \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \|\overline{G}_N^{-1/2} Q_N(\theta)(\theta-\overline{\theta})\|_2 \quad \text{(using Eq.(6.41))} \\
&= \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \|\overline{G}_N^{-1/2} Q_N(\theta)(\theta-\overline{\theta})\|_2 \times \frac{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \\
&\geq \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \frac{(\theta-\overline{\theta})^\top Q_N(\theta)(\theta-\overline{\theta})}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \quad \text{(using the Cauchy Schwarz's inequality)} \\
&= \delta \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \frac{(\theta-\overline{\theta})^\top \overline{G}_N^{1/2}}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \overline{G}_N^{-1/2} Q_N(\theta) \overline{G}_N^{-1/2} \frac{\overline{G}_N^{1/2}(\theta-\overline{\theta})}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \\
&\geq \delta \inf_{\|e\|_2=1,\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} e^\top \overline{G}_N^{-1/2} Q_N(\theta) \overline{G}_N^{-1/2} e \\
&= \delta \inf_{\|e\|_2=1,\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} e^\top \overline{G}_N^{-1/2} \int_0^1 G_N(t\overline{\theta}+(1-t)\theta)dt \overline{G}_N^{-1/2} e \\
&= \delta \inf_{\|e\|_2=1,\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \int_0^1 \left(e^\top \overline{G}_N^{-1/2} G_N(t\overline{\theta}+(1-t)\theta)\overline{G}_N^{-1/2} e\right) dt \\
&\geq \delta \inf_{\|e\|_2=1,\theta\in\mathcal{N}_N(\overline{\theta},\delta)} e^\top \overline{G}_N^{-1/2} G_N(\theta) \overline{G}_N^{-1/2} e \\
&\geq \delta \left\{ \inf_{\|e\|_2=1} e^\top \overline{G}_N^{-1/2} G_N(\overline{\theta}) \overline{G}_N^{-1/2} e - \mathcal{C}\frac{\delta}{N^{1/2}} \right\} =: \mathcal{I}_N(\delta,\overline{\theta}),
\end{aligned}
\tag{6.42}
$$

where in the penultimate inequality we used that $\overline{\theta} \in \mathcal{N}_N(\overline{\theta},\delta)$ and the convexity of $\mathcal{N}_N(\overline{\theta},\delta)$. In the last inequality, we used Lemma 6.16 whose proof is postponed to Section 6.8.6.

Lemma 6.16. *Let us consider some $\delta > 0$. Then for any $N \in \mathbb{N}$ and for any unit vector $u \in \mathbb{R}^s$, it holds*

$$
\sup_{\theta\in\mathcal{N}_N(\overline{\theta},\delta)} |u^\top \overline{G}_N^{-1/2}(G_N(\theta)-G_N(\overline{\theta}))\overline{G}_N^{-1/2} u| \leq \mathcal{C}\frac{\delta}{N^{1/2}},
$$

where $\mathcal{N}_N(\overline{\theta},\delta) = \{\theta \in \mathbb{R}^s \ : \ \|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2 \leq \delta\}$ and where $\mathcal{C}$ is a constant that only depends on the quantities $s, K, c, \overline{\sigma}^2_{\min}$ (that do not depend on N).

To lower bound uniformly in N the term $\mathcal{I}_N(\delta,\overline{\theta})$, we notice that

$$
\begin{aligned}
&\inf_{\|e\|_2=1} e^\top \overline{G}_N^{-1/2} G_N(\overline{\theta}) \overline{G}_N^{-1/2} e \\
&= \inf_{\|e\|_2=1} \frac{e^\top \overline{G}_N^{-1/2}}{\|\overline{G}_N^{-1/2} e\|_2} G_N(\overline{\theta}) \frac{\overline{G}_N^{-1/2} e}{\|\overline{G}_N^{-1/2} e\|_2} \|\overline{G}_N^{-1/2} e\|_2^2 \\
&\geq \lambda_{\min}(G_N(\overline{\theta})) \inf_{\|e\|_2=1} \|\overline{G}_N^{-1/2} e\|_2^2 \\
&\geq \lambda_{\min}(G_N(\overline{\theta}))\, \lambda_{\min}(\overline{G}_N^{-1}) \\
&\geq \left(\overline{\sigma}^2_{\min} cN\right) \times \left(4C^{-1}N^{-1}\right) \\
&\geq 4C^{-1} c \overline{\sigma}^2_{\min},
\end{aligned}
$$

where we used that for any $i \in [N]$, $\sigma'(\mathbf{x}_{i,M}\overline{\theta}) \geq \overline{\sigma}^2_{\min}$. Let us denote $N_\delta := \lceil \left(\frac{\mathcal{C}\delta}{2C^{-1}c\overline{\sigma}^2_{\min}}\right)^2 \rceil$ so that for any $N \geq N_\delta$ it holds

$$
\mathcal{I}_N(\delta,\overline{\theta}) \geq \delta 2C^{-1} c \overline{\sigma}^2_{\min}.
$$

Using Markov's inequality, we get that for any $N \geq N_\delta$,

$$\begin{aligned}
&\overline{\mathbb{P}}_{\theta^*}(\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 \geq \mathcal{I}_N(\delta, \overline{\theta})) \\
&\leq (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}(\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2^2) \\
&\leq (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}((Y - \overline{\pi}^{\theta^*})^\top \mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*})) \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}(\mathrm{Tr}\left[(Y - \overline{\pi}^{\theta^*})^\top \mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*})\right]) \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}(\mathrm{Tr}\left[\mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*})(Y - \overline{\pi}^{\theta^*})^\top\right]) \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \mathrm{Tr}\left[\mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*}\right] \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \mathrm{Tr}\left[\overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M\right].
\end{aligned}$$

Hence, it holds for any $N \geq N_\delta$,

$$\begin{aligned}
&\overline{\mathbb{P}}_{\theta^*}(\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 \geq \mathcal{I}_N(\delta, \overline{\theta})) \\
&\leq \frac{\mathrm{Tr}\left[\overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M\right]}{\mathcal{I}_N(\delta, \overline{\theta})^2} \\
&< \frac{\mathfrak{K}}{\delta^2 (2C^{-1} c \overline{\sigma}_{\min}^2)^2} \\
&\leq \epsilon,
\end{aligned} \tag{6.43}$$

where the last inequality comes from the choice of δ (see Eq.(6.38)). From Eq.(6.42) and Eq.(6.43), we deduce that for any $N \geq N_\delta$, it holds

$$\overline{\mathbb{P}}_{\theta^*}(E_N) \geq 1 - \epsilon,$$

where

$$E_N := \left\{ \|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 < \inf_{\theta \in \partial \mathcal{N}_N(\overline{\theta}, \delta)} \|F(\theta)\|_2 \right\}.$$

Hence, on the event E_N, we define $\widehat{\theta} = F^{-1}(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M)))$ which means by definition of F that $\widehat{\theta}$ is the conditional MLE, namely

$$L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0.$$

Third and final step. In the previous step, we proved that for N large enough, the MLE exists and is contained in an ellipsoid centered at $\overline{\theta}$ with vanishing volume with high probability. Now we show how using this result to turn the CLT on $L_N(\overline{\theta}, (Y, \mathbf{X}_M))$ from Theorem 6.8 into a CLT for $\widehat{\theta}$.
We consider $N \geq N_\delta$ and we work on the event E_N of the previous step. Since $L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0$ by definition of $\widehat{\theta}$, we get that

$$\begin{aligned}
L_N(\overline{\theta}, (Y, \mathbf{X}_M)) &= L_N(\overline{\theta}, (Y, \mathbf{X}_M)) - L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) \\
&= \mathbf{X}_M^\top (\pi^{\overline{\theta}} - \pi^{\widehat{\theta}}) \\
&= \underbrace{\int_0^1 G_N(t\overline{\theta} + (1-t)\widehat{\theta}) dt}_{= Q_N(\widehat{\theta})} (\overline{\theta} - \widehat{\theta}),
\end{aligned}$$

where we used that the Jacobian of the map $\theta \mapsto \mathbf{X}_M^\top \pi^\theta = \mathbf{X}_M \sigma(\mathbf{X}_M \theta)$ is $\mathbf{X}_M \mathrm{Diag}(\sigma'(\mathbf{X}_M \theta)) \mathbf{X}_M = G_N(\theta)$. From the Portmanteau Theorem [cf. Van der Vaart, 2000, Lemma 2.2]), we know that a sequence of $\mathbb{R}^s$-valued random vectors $(X_n)_n$ converges weakly to a random vector X if and only if for any Lipschitz and bounded function $h : \mathbb{R}^s \to \mathbb{R}$ it holds

$$\mathbb{E}h(X_n) \underset{n \to \infty}{\to} \mathbb{E}h(X).$$

Hence, we consider a Lipschitz and bounded function $h : \mathbb{R}^s \to \mathbb{R}$. We denote by $L_h > 0$ the Lipschitz constant of h. It holds for any $N \geq N_\delta$,

$$
\begin{aligned}
&|\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}\big[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\big]| \\
&= |\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta}))]| \\
&\leq |\overline{\mathbb{E}}_{\theta^*}\Big[\mathbb{1}_{E_N}\Big\{h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta})) - h\big(\overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta})\big)\Big\}\Big]| + 2\|h\|_\infty \overline{\mathbb{P}}_{\theta^*}(E_N^c) \\
&\leq \overline{\mathbb{E}}_{\theta^*}\big[L_h \mathbb{1}_{E_N} \|\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}) - \overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta})\|_2\big] + 2\|h\|_\infty \epsilon \\
&\leq L_h \overline{\mathbb{E}}_{\theta^*}\big[\mathbb{1}_{E_N} \|\overline{G}_N^{-1/2}(G_N(\overline{\theta}) - Q_N(\widehat{\theta}))\overline{G}_N^{-1/2}\| \|\overline{G}_N^{1/2}(\overline{\theta} - \widehat{\theta})\|_2\big] + 2\|h\|_\infty \epsilon \\
&\leq L_h \delta \sup_{\theta \in \mathcal{N}_N(\overline{\theta}, \delta)} \|\overline{G}_N^{-1/2}(G_N(\overline{\theta}) - Q_N(\theta))\overline{G}_N^{-1/2}\| + 2\|h\|_\infty \epsilon, \qquad (6.44)
\end{aligned}
$$

where we used that on the event E_N, $\widehat{\theta} \in \mathcal{N}_N(\overline{\theta}, \delta)$, i.e. $\|\overline{G}_N^{1/2}(\overline{\theta} - \widehat{\theta})\|_2 \leq \delta$. Moreover, for any $\theta' \in \mathcal{N}_N(\overline{\theta}, \delta)$ we have,

$$
\begin{aligned}
&\|\overline{G}_N^{-1/2}(G_N(\overline{\theta}) - Q_N(\theta'))\overline{G}_N^{-1/2}\| \\
&= \sup_{\|u\|_2 = 1} |u^\top \overline{G}_N^{-1/2}(G_N(\overline{\theta}) - Q_N(\theta'))\overline{G}_N^{-1/2} u| \\
&\leq \sup_{\|u\|_2 = 1} \int_0^1 \Big|u^\top \overline{G}_N^{-1/2}(G_N(\overline{\theta}) - G_N(t\overline{\theta} + (1-t)\theta'))\overline{G}_N^{-1/2} u\Big| \, dt \\
&\leq \sup_{\|u\|_2 = 1} \sup_{\theta \in \mathcal{N}_N(\overline{\theta}, \delta)} |u^\top \overline{G}_N^{-1/2}(G_N(\overline{\theta}) - G_N(\theta))\overline{G}_N^{-1/2} u| \\
&\leq \mathcal{C} \frac{\delta}{N^{1/2}}, \qquad (6.45)
\end{aligned}
$$

where in the penultimate inequality we used the convexity of the set $\mathcal{N}_N(\overline{\theta}, \delta)$ and in the last inequality we used Lemma 6.16 (which is proved in Section 6.8.6). Using Eq.(6.44) and Eq.(6.45), we deduce that for $G \sim \mathcal{N}(0, \mathrm{Id}_s)$ we have

$$
\begin{aligned}
&|\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \mathbb{E}[h(G)]| \\
&\leq |\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}\big[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\big]| \\
&\quad + |\overline{\mathbb{E}}_{\theta^*}\big[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\big] - \mathbb{E}[h(G)]| \\
&\leq L_h \delta \mathcal{C} \frac{\delta}{N^{1/2}} + 2\|h\|_\infty \epsilon + |\overline{\mathbb{E}}_{\theta^*}\big[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\big] - \mathbb{E}[h(G)]|. \qquad (6.46)
\end{aligned}
$$

The CLT from Theorem 6.8 states that

$$
\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M)) \xrightarrow[N \to \infty]{(d)} \mathcal{N}(0, \mathrm{Id}_s),
$$

which means by the Portmanteau Theorem [cf. Van der Vaart, 2000, Lemma 2.2]) that

$$
|\overline{\mathbb{E}}_{\theta^*}\big[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\big] - \mathbb{E}[h(G)]| \underset{N \to +\infty}{\to} 0.
$$

We deduce that for any $\epsilon > 0$ and for any Lipschitz and bounded function $h : \mathbb{R}^s \to \mathbb{R}$, one can choose N large enough to ensure that the right hand side of Eq.(6.46) is smaller than $4\|h\|_\infty \epsilon$. Note that this is true since the constant δ does not depend on N. This concludes the proof thanks to the Portmanteau Theorem.

6.8.6 Proof of Lemma 6.16

Let us first recall that $G_N(\overline{\theta}) = \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\overline{\theta}))\mathbf{X}_M$ and that $\mathbf{X}_M^\top = [\mathbf{w}_1 \mid \mathbf{w}_2 \mid \ldots \mid \mathbf{w}_N]$, where $\mathbf{w}_i = \mathbf{x}_{i,M} \in \mathbb{R}^s$. Let us consider some $\theta \in \mathcal{N}_N(\overline{\theta}, \delta)$. We have that

$$\begin{aligned} G_N(\theta) - G_N(\overline{\theta}) &= \sum_{i=1}^N \mathbf{w}_i \left[\sigma'(\mathbf{w}_i^\top\theta) - \sigma'(\mathbf{w}_i^\top\overline{\theta})\right] \mathbf{w}_i^\top \\ &= \sum_{i=1}^N \mathbf{w}_i \underbrace{\int_0^1 \sigma''(t\mathbf{w}_i^\top\theta + (1-t)\mathbf{w}_i^\top\overline{\theta})dt}_{=:H_i} \mathbf{w}_i^\top(\theta-\overline{\theta})\mathbf{w}_i^\top. \end{aligned} \tag{6.47}$$

We get using Eq.(6.47) that for any unit vector $u \in \mathbb{R}^s$,

$$\begin{aligned} &|u^\top \overline{G}_N^{-1/2}(G_N(\theta) - G_N(\overline{\theta}))\overline{G}_N^{-1/2}u| \\ &= \left|\sum_{i=1}^N u^\top \overline{G}_N^{-1/2}\mathbf{w}_i H_i \mathbf{w}_i^\top(\theta-\overline{\theta})\mathbf{w}_i^\top \overline{G}_N^{-1/2}u\right| \\ &= \left|\sum_{i=1}^N \mathbf{w}_i^\top(\theta-\overline{\theta}) \times u^\top \overline{G}_N^{-1/2}\mathbf{w}_i H_i \mathbf{w}_i^\top \overline{G}_N^{-1/2}u\right| \\ &= \left|\sum_{i=1}^N \mathbf{w}_i^\top(\theta-\overline{\theta}) \times H_i |\mathbf{w}_i^\top \overline{G}_N^{-1/2}u|^2\right| \\ &\leq \max_{1\leq j\leq N} |\mathbf{w}_j^\top(\theta-\overline{\theta})| \sum_{i=1}^N |H_i||\mathbf{w}_i^\top \overline{G}_N^{-1/2}u|^2 \\ &= \max_{1\leq j\leq N} |\mathbf{w}_j^\top(\theta-\overline{\theta})| \; \|\mathbf{H}^{1/2}\mathbf{X}_M^\top \overline{G}_N^{-1/2}u\|_2^2, \end{aligned} \tag{6.48}$$

where $\mathbf{H}^{1/2} := \mathrm{Diag}((|H_i|^{1/2})_{i\in[N]})$. The proof is concluded by upper-bounding both terms involved in the product of the right hand side of Eq.(6.48). Using the assumption of the design matrix presented in Section 6.5.1 and recalling that $\theta \in \mathcal{N}_N(\overline{\theta}, \delta)$, we have

$$\begin{aligned} \max_{1\leq j\leq N} |\mathbf{w}_j^\top(\theta-\overline{\theta})| &\leq \max_{1\leq j\leq N} \|\overline{G}_N^{-1/2}\mathbf{w}_j\|_2 \underbrace{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}_{\leq\delta} \\ &= \delta K\sqrt{(\overline{\sigma}_{\min}^2 c)^{-1} s}N^{-1/2}, \end{aligned}$$

where we used that $\|\overline{G}_N^{-1/2}\|^2 = \|\overline{G}_N^{-1}\| \leq (c\overline{\sigma}_{\min}^2 N)^{-1}$ and that for any $i \in [N]$, $\|\mathbf{w}_i\|_2^2 \leq sK^2$. Since $|H_i| \leq 1$ for any $i \in [N]$,

$$\begin{aligned} \|\mathbf{H}^{1/2}\mathbf{X}_M^\top \overline{G}_N^{-1/2}u\|_2^2 &\leq \|\mathbf{X}_M^\top \overline{G}_N^{-1/2}u\|_2^2 \\ &= \sum_{i=1}^N (\mathbf{w}_i^\top \overline{G}_N^{-1/2}u)^2 \\ &\leq \sum_{i=1}^N \|\overline{G}_N^{-1/2}\mathbf{w}_i\|_2^2 \leq (\overline{\sigma}_{\min}^2 c)^{-1} sK^2, \end{aligned}$$

where in the penultimate inequality we used Cauchy-Schwarz inequality.

6.8.7 Proof of Proposition 6.10

For any $N \in \mathbb{N}$, let us denote

$$\mathcal{E}_N := \{Z \in \{0,1\}^N \mid \mathbf{X}_M^\top Z \in \mathrm{Im}(\Xi)\}. \tag{6.49}$$

In order to clarify the notations of this proof, let us stress that we denote in the following by $\overline{\mathbb{P}}_{\theta_0^*}$ the distribution of Y, $\mathbb{P}_1$ the distribution of the sequence $(Y^{(t)})_{t\geq 1}$ and $\mathbb{P}_2$ the distribution of $(Z^{(t)})_{t\geq 1}$. Let us consider some $\epsilon > 0$.

Step 1: $\mathbb{P}_1$ almost sure convergences.
From Proposition 6.7, we know that under the null $\mathbb{H}_0$

$$\frac{\sum_{t=1}^T Y^{(t)} \mathbb{P}_{\theta_0^*}(Y^{(t)})}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})} \underset{T\to\infty}{\to} \overline{\mathbb{E}}_{\theta_0^*}[Y] = \overline{\pi}^{\theta_0^*} \quad \mathbb{P}_1 - \text{almost surely.} \tag{6.50}$$

Since $\widetilde{\pi}^{\theta_0^*} \underset{T\to\infty}{\to} \overline{\pi}^{\theta_0^*}$ $\mathbb{P}_1$-a.s., we know that $\mathbb{P}_1$-a.s, there exists some $T_1 \in \mathbb{N}$ such that for any $T \geq T_1$ it holds

$$\|\widetilde{\pi}^{\theta_0^*} \odot (1 - \widetilde{\pi}^{\theta_0^*}) - \overline{\pi}^{\theta_0^*} \odot (1 - \overline{\pi}^{\theta_0^*})\|_\infty < \epsilon,$$

and since $(\overline{\sigma}^{\theta_0^*})^2 \geq (\sigma_{\min})^2 > 0$, we get by continuity of the inverse of a matrix that $\mathbb{P}_1$-a.s, there exists some $T_2 \in \mathbb{N}$ such that for any $T \geq T_2$, it holds

$$\|\widetilde{G}_N^{-1} - \overline{G}_N^{-1}\| < \epsilon^2,$$

where we recall that

$$\widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}(\widetilde{\pi}^{\theta_0^*} \odot (1 - \widetilde{\pi}^{\theta_0^*}))\mathbf{X}_M,$$

and

$$\overline{G}_N = \mathbf{X}_M^\top \mathrm{Diag}(\overline{\pi}^{\theta_0^*} \odot (1 - \overline{\pi}^{\theta_0^*}))\mathbf{X}_M.$$

From Eq.(6.50) and by continuity of the map Ψ, we get that $\mathbb{P}_1$-a.s. $\widetilde{\theta} = \Psi(\mathbf{X}_M^\top \widetilde{\pi}^{\theta_0^*}) \underset{T\to\infty}{\to} \Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}) = \overline{\theta}(\theta_0^*)$ (see Eq.(6.30)). Hence, $\mathbb{P}_1$-a.s, there exists some $T_3 \in \mathbb{N}$ such that for any $T \geq T_3$, it holds

$$\|\widetilde{\theta} - \overline{\theta}\|_2 \leq \epsilon.$$

Note that we left the dependence of $\widetilde{\pi}^{\theta_0^*}$ and $\widetilde{\theta}$ on T implicit.

Step 2: Comparing $\widetilde{W}_N$ and W_N.
It holds for any $Z \in \mathcal{E}_N$,

$$\begin{aligned}
&\left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \widetilde{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right| \\
&\leq \left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right| \\
&\quad + \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\overline{\theta} - \widetilde{\theta} \right) \right\|_2 \\
&\leq \|\widetilde{G}_N^{-1/2} - \overline{G}_N^{-1/2}\| \|G_N(\widetilde{\theta})\| \left\| \Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right\|_2 \\
&\quad + \|\overline{G}_N^{-1/2}\| \|G_N(\widetilde{\theta}) - G_N(\overline{\theta})\| \left\| \Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right\|_2 + \left\| \mathbf{X}_M^\top \mathbf{X}_M \right\| \|\overline{\theta} - \widetilde{\theta}\|_2.
\end{aligned}$$

Using the Powers–Størmer inequality [cf. Powers and Størmer, 1970, Lemma 4.1] and denoting $\|M\|_1$ the Schatten 1-norm of any matrix M, it holds

$$\|\widetilde{G}_N^{-1/2} - \overline{G}_N^{-1/2}\|^2 \leq \|\widetilde{G}_N^{-1/2} - \overline{G}_N^{-1/2}\|_F^2 \leq \|\widetilde{G}_N^{-1} - \overline{G}_N^{-1}\|_1 \leq 2s \|\widetilde{G}_N^{-1} - \overline{G}_N^{-1}\|,$$

where in the last inequality we used that $\widetilde{G}_N$ and $\overline{G}_N$ have rank at most s. Hence, $\mathbb{P}_1$-a.s, for any $T \geq T_N(\epsilon) := \max(T_1, T_2, T_3)$ it holds

$$\begin{aligned}
&\left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \widetilde{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right| \\
&\leq \left\| \Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right\|_2 \left\{ \epsilon 2sCN + (c(\overline{\sigma}_{\min})^2 N)^{-1/2} CN\epsilon \right\} + CN\epsilon =: \mathcal{C}_N(Z, \epsilon).
\end{aligned}$$

We get that $\mathbb{P}_1$-a.s, for any $T \geq T_N(\epsilon)$ it holds

$$\sup_{Z\in\mathcal{E}_N} \left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \widetilde{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right| \leq \sup_{Z\in\mathcal{E}_N} \mathcal{C}_N(Z,\epsilon) =: \mathcal{C}_N(\epsilon).$$

Step 3: Conclusion.
Let us consider some $\eta \in (0, 1-\alpha)$. Since $\mathcal{C}_N(\epsilon)$ goes to 0 as $\epsilon \to 0$, we deduce that we can choose ϵ small enough such that $\mathbb{P}_1$-a.s., for any $T \geq T_N(\epsilon)$ it holds

$$\forall Z \in \mathcal{E}_N, \quad \mathbb{1}_{Z\in\widetilde{W}_N} \leq \mathbb{1}_{Z\in W_N(\alpha+\eta)}, \tag{6.51}$$

where

$$W_N(\alpha+\eta) := \left\{ Z \in \{0,1\}^N \;\middle|\; \begin{array}{l} \diamond\ \mathbf{X}_M^\top Z \in \mathrm{Im}(\Xi) \\ \diamond\ \left\| [\overline{G}_N]^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2^2 > \chi^2_{s,1-\alpha-\eta} \end{array} \right\},$$

Recalling the definition of $\mathcal{E}_N$ from Eq.(6.49) and using the definitions of $W_N(\alpha+\eta)$ and $\widetilde{W}_N$, it also holds trivially

$$\forall Z \in \{0,1\}^N \backslash \mathcal{E}_N, \quad 0 = \mathbb{1}_{Z\in\widetilde{W}_N} \leq \mathbb{1}_{Z\in W_N(\alpha+\eta)} = 0. \tag{6.52}$$

Using both Eq.(6.51) and Eq.(6.52), we deduce that

$$\forall Z \in \{0,1\}^N, \quad \mathbb{1}_{Z\in\widetilde{W}_N} \leq \mathbb{1}_{Z\in W_N(\alpha+\eta)},$$

and we then get that $\mathbb{P}_1$-a.s., for any $T \geq T_N(\epsilon)$, we have

$$\zeta_{N,T} = \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)}) \mathbb{1}_{Z^{(t)}\in\widetilde{W}_N}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)})} \leq \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)}) \mathbb{1}_{Z^{(t)}\in W_N(\alpha+\eta)}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)})}.$$

The right hand side of the previous inequality converges $\mathbb{P}_2$-a.s. to $\overline{\mathbb{P}}_{\theta_0^*}(Y \in W_N(\alpha+\eta))$ as $T \to +\infty$ thanks to Proposition 6.7. Since from Theorem 6.9 it holds,

$$\limsup_{N\to+\infty} \overline{\mathbb{P}}_{\theta_0^*}(Y \in W_N(\alpha+\eta)) \leq \alpha+\eta,$$

we get that for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any $N \geq N_0$ it holds,

$$\mathbb{P}\Big(\bigcup_{T_N\in\mathbb{N}} \bigcap_{T\geq T_N} \{\zeta_{N,T} \leq \alpha+\epsilon\} \Big) = 1.$$

6.8.8 Proof of Proposition 6.11

Let us denote $\mathcal{M} : \theta \in \mathbb{R}^s \mapsto \mathbf{X}_M^\top \overline{\pi}^\theta$. Since for any $z \in \{0,1\}^N$, $\mathbb{P}_\theta(z) = \exp(-\mathcal{L}_N(\theta,(z,\mathbf{X}_M)))$, we get $\nabla_\theta \mathbb{P}_\theta(z) = -L_N(\theta,(z,\mathbf{X}_M))\mathbb{P}_\theta(z)$. Recalling that $\overline{\pi}^\theta = \overline{\mathbb{E}}_\theta[Y]$, we have for any $k \in [s]$,

$$\begin{aligned} \frac{\partial \overline{\pi}^\theta}{\partial \theta_k} &= \left(\sum_{w\in E_M} \mathbb{P}_\theta(w) \right)^{-2} \sum_{w,z\in E_M} \mathbb{P}_\theta(z)\mathbb{P}_\theta(w) z \left\{ L_N(\theta,(w,\mathbf{X}_M)) - L_N(\theta,(z,\mathbf{X}_M)) \right\}_k \\ &= \overline{\mathbb{E}}_\theta \left[Z \left\{ L_N(\theta,(W,\mathbf{X}_M)) - L_N(\theta,(Z,\mathbf{X}_M)) \right\}_k \right] \\ &= \overline{\mathbb{E}}_\theta \left[Z \left\{ \mathbf{X}_M^\top (Z-W) \right\}_k \right] \\ &= \overline{\Gamma}^\theta \mathbf{X}_{:,M[k]}, \end{aligned} \tag{6.53}$$

where Z and W are independent random vectors valued in $\{0,1\}^N$ and distributed according to $\overline{\mathbb{P}}_\theta$. Note that we used that for any $W \in \{0,1\}^N$, it holds

$$L_N(\theta,(W,\mathbf{X}_M)) = \mathbf{X}_M^\top(\sigma(\mathbf{X}_M\theta) - W).$$

Hence it holds

$$\forall \theta \in \mathbb{R}^s, \quad \nabla \mathcal{M}(\theta) = \mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M.$$

Suppose that we are able to compute an estimate $\theta^\star \in \mathbb{B}_p(0,R)$ of θ^*. Using that $\theta^* \in \mathbb{B}_p(0,R)$ and that

$$\inf_{\theta \in \mathbb{B}_p(0,R)} \lambda_{\min}\left(\nabla \mathcal{M}(\theta)\right) \geq \kappa \lambda_{\min}\left(\mathbf{X}_M^\top \mathbf{X}_M\right) \geq c\kappa N,$$

it holds

$$\begin{aligned}
\|\mathcal{M}(\theta^\star) - \mathcal{M}(\theta^*)\|_2^2 &= \| \int_0^1 \nabla \mathcal{M}(t\theta^\star + (1-t)\theta^*)(\theta^\star - \theta^*) dt\|_2^2 \\
&= (\theta^\star - \theta^*)^\top \left\{ \int_0^1 \nabla \mathcal{M}(t\theta^\star + (1-t)\theta^*) dt \right\}^2 (\theta^\star - \theta^*) \\
&\geq \|\theta^\star - \theta^*\|_2^2 \inf_{\theta \in \mathbb{B}_p(0,R)} \lambda_{\min}(\nabla \mathcal{M}(\theta))^2 \\
&\geq (c\kappa N)^2 \|\theta^\star - \theta^*\|_2^2.
\end{aligned}$$

Noticing further that

$$\sup_{\theta \in \mathbb{R}^s} \|\nabla \Psi^{-1}(\theta)\| = \sup_{\theta \in \mathbb{R}^s} \|\mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M \theta))\mathbf{X}_M\| \leq \frac{1}{4} CN,$$

we get

$$\begin{aligned}
\|\theta^* - \theta^\star\|_2 &\leq (\kappa c N)^{-1} \|\mathbf{X}_M^\top \overline{\pi}^{\theta^\star} - \mathbf{X}_M^\top \overline{\pi}^{\theta^*}\|_2 \\
&= (\kappa c N)^{-1} \|\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^\star)} - \mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^*)}\|_2 \quad \text{(using Eq.(6.30))} \\
&\leq (\kappa c N)^{-1} \sup_{\theta \in \mathbb{R}^s} \|\nabla \Psi^{-1}(\theta)\| \|\Psi\left(\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^\star)}\right) - \Psi\left(\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^*)}\right)\|_2 \\
&\leq C\,(\kappa c)^{-1} \|\Psi\left(\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^\star)}\right) - \Psi\left(\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^*)}\right)\|_2 \\
&= C\,(\kappa c)^{-1} \|\overline{\theta}(\theta^\star) - \overline{\theta}(\theta^*)\|_2 \\
&\leq C\,(\kappa c)^{-1} \left[\|\overline{\theta}(\theta^\star) - \widehat{\theta}\|_2 + \|\widehat{\theta} - \overline{\theta}(\theta^*)\|_2\right],
\end{aligned}$$

where we used that $\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^*)} = \mathbf{X}_M^\top \sigma\big(\mathbf{X}_M \overline{\theta}(\theta^*)\big) = \Xi\big(\overline{\theta}(\theta^*)\big) \in \mathrm{Im}(\Xi)$ and thus $\Psi(\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^*)})$ is well-defined. Similarly, we have that $\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta^\star)} \in \mathrm{Im}(\Xi)$. Since Theorem 6.9 gives that

$$\overline{\mathbb{P}}_{\theta^*}\left(\|V_N(\theta^*)(\widehat{\theta} - \overline{\theta})\|_2^2 \leq \chi^2_{s,1-\alpha}\right) \underset{N \to +\infty}{\to} 1 - \alpha,$$

with $V_N(\theta^*) := [\overline{G}_N(\theta^*)]^{-1/2} G_N(\overline{\theta}(\theta^*))$, we deduce (using the assumption of the design matrix from Section 6.5.1) that the event

$$\|\widehat{\theta} - \overline{\theta}(\theta^*)\|_2 \leq \|[V_N(\theta^*)]^{-1}\| \|V_N(\theta^*)(\widehat{\theta} - \overline{\theta})\|_2 \leq \|(\sigma^{\overline{\theta}})^{-2}\|_\infty c^{-1} (N/C)^{-1/2} \sqrt{\chi^2_{s,1-\alpha}},$$

holds with probability tending to $1-\alpha$ as $N \to +\infty$. Note that we used that

$$\|G_N(\overline{\theta}(\theta^*))^{-1}\| \leq (cN)^{-1} \|(\sigma^{\overline{\theta}})^{-2}\|_\infty,$$

and that

$$\|[\overline{G}_N(\theta^*)]^{1/2}\| \leq (CN)^{1/2}.$$

Hence we obtain an asymptotic confidence region for θ^* of level $1-\alpha$.

6.8.9 Proof of Proposition 6.13

Let us denote $\mathcal{R} : \pi \in (0,1)^N \mapsto \overline{\pi}^{\pi}$. It holds for any $i \in [N]$,

$$\begin{aligned}\frac{\partial \overline{\pi}^{\pi}}{\partial \pi_i} &= \left(\sum_{w \in E_M} \mathbb{P}_\pi(w)\right)^{-2} \sum_{w,z \in E_M} \mathbb{P}_\pi(z)\mathbb{P}_\pi(w) z \left\{z - w\right\}_i \left(\pi_i(1-\pi_i)\right)^{-1} \\ &= \overline{\mathbb{E}}_\pi \left[Z(Z-W)_i^\top\right] \left(\pi_i(1-\pi_i)\right)^{-1},\end{aligned}$$

where Z and W are independent random vectors valued in $\{0,1\}^N$ and distributed according to $\overline{\mathbb{P}}_\pi$. Hence it holds

$$\forall \pi \in (0,1)^N, \quad \nabla \mathcal{R}(\pi) = \overline{\Gamma}^{\pi} \mathrm{Diag}(\pi \odot (1-\pi))^{-1}.$$

Suppose that we are able to compute an estimate $\pi^\bigstar \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R)$ of π^*. Then since it holds for any $v \in \mathbb{R}^N$,

$$\inf_{\pi \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R)} \|\nabla \mathcal{R}(\pi) v\|_2 \geq 4\kappa \|v\|_2,$$

we get that

$$\begin{aligned}\|\mathcal{R}(\pi^\bigstar) - \mathcal{R}(\pi^*)\|_2 &= \| \int_0^1 \nabla \mathcal{R}(t\pi^\bigstar + (1-t)\pi^*)(\pi^\bigstar - \pi^*) dt \|_2 \\ &\geq 4\kappa \|\pi^\bigstar - \pi^*\|_2.\end{aligned}$$

Hence we have that

$$\begin{aligned}\|\pi^* - \pi^\bigstar\|_2 &\leq (4\kappa)^{-1} \|\overline{\pi}^{\pi^\bigstar} - \overline{\pi}^{\pi^*}\|_2 \\ &\leq (4\kappa)^{-1} \big\{ \|\mathrm{Proj}_{\mathbf{X}_M}(\overline{\pi}^{\pi^\bigstar} - Y)\|_2 + \|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^*})\|_2 \\ &\quad + \|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^\bigstar} - \overline{\pi}^{\pi^*})\|_2 \big\}.\end{aligned}$$

Since Theorem 6.8 gives that

$$\overline{\mathbb{P}}_{\pi^*} \left(\|[\overline{G}_N(\pi^*)]^{-1/2} (\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi^*})\|_2^2 \leq \chi^2_{s,1-\alpha} \right) \underset{N \to +\infty}{\to} 1 - \alpha,$$

we deduce that the event

$$\begin{aligned}\|\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi^*}\|_2 &\leq \|[\overline{G}_N(\pi^*)]^{1/2}\| \|[\overline{G}_N(\pi^*)]^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*})\|_2 \\ &\leq (CN)^{1/2} \sqrt{\chi^2_{s,1-\alpha}},\end{aligned}$$

holds with probability tending to $1-\alpha$ as $N \to +\infty$. Noticing further that for any vector $v \in \mathbb{R}^N$,

$$\|\mathrm{Proj}_{\mathbf{X}_M} v\|_2 \leq \|\mathbf{X}_M \left(\mathbf{X}_M^\top \mathbf{X}_M\right)^{-1}\| \times \|\mathbf{X}_M^\top v\|_2 \leq (CN)^{1/2} (cN)^{-1} \|\mathbf{X}_M^\top v\|_2,$$

we get that for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any $N \geq N_0$, it holds with at least $1 - \alpha - \epsilon$,

$$\begin{aligned}\|\pi^* - \pi^\bigstar\|_2 \leq (4\kappa)^{-1} \big\{ &\|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^\bigstar})\|_2 + Cc^{-1} \sqrt{\chi^2_{s,1-\alpha}} \\ &+ \|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^\bigstar} - \overline{\pi}^{\pi^*})\|_2 \big\}.\end{aligned}$$

Hence we obtain an asymptotic confidence region for π^* of level $1-\alpha$.

6.9 Inference conditional on the signs

6.9.1 Leftover Fisher information

As highlighted in Fithian et al. [2014], conducting inference conditional on some random variable prevents the use of this variable as evidence against a hypothesis. Selective inference should be understood as partitioning the observed information in two sets: the one used to select the model and the one used

to make inference. This communicating vessels principle is illustrated with the following inclusions borrowed from Fithian et al. [2014].

$$\mathcal{F}_0 \underbrace{\subset}_{\text{used for selection}} \mathcal{F}(\mathbb{1}_{Y\in\mathcal{M}}) \underbrace{\subset}_{\text{used for inference}} \mathcal{F}(Y).$$

Typically, let us assume that we condition on both the selected support $\widehat{M}(Y) = M$ and the observed vector of signs $\widehat{S}_M(Y) = S_M \in \{0,1\}^{|M|}$, meaning that $\mathcal{M} = E_M^{S_M}$ (cf. Eq.(6.15)). Even if the vector of signs S_M is surprising under $\mathbb{H}_0$, we will not reject unless we are surprised anew by observing the response variable Y. Stated otherwise, when we condition on both the selected support and the vector of signs, we cannot take advantage of the possible unbalanced probability distribution of the vector of signs $\widehat{S}_M(Y)$ conditionally on E_M. Hence, conditioning on a finer σ-algebra results in some information loss. Fithian et al. [2014] explain that we can actually quantify this waste of information. The Hessian of the log-likelihood can be decomposed as

$$\nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, E_M) = \nabla^2_\vartheta \mathcal{L}_N(\vartheta, \widehat{S}_M(Y) \,|\, E_M) + \nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, \{E_M, \widehat{S}_M(Y)\}). \tag{6.54}$$

For any σ-algebra $\mathcal{F} \subseteq \sigma(Y)$, we consider the conditional expectation

$$\mathcal{I}_{Y\,|\,\mathcal{F}}(\vartheta) := -\mathbb{E}\left[\nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, \mathcal{F}) \,|\, \mathcal{F}\right].$$

The *leftover Fisher information* after selection at $\widehat{S}_M(Y)$ is defined by $\mathcal{I}_{Y\,|\,\{E_M, \widehat{S}_M(Y)\}}(\vartheta)$. Taking expectation in both sides of Eq.(6.54) leads to

$$\begin{aligned}\mathbb{E}\left[\mathcal{I}_{Y\,|\,\{E_M,\widehat{S}_M(Y)\}}(\vartheta)\right] &= \mathbb{E}\,\mathcal{I}_{Y\,|\,E_M}(\vartheta) - \mathbb{E}\,\mathcal{I}_{\widehat{S}_M(Y)\,|\,E_M}(\vartheta) \\ &\preceq \mathbb{E}\,\mathcal{I}_{Y\,|\,E_M}(\vartheta),\end{aligned}$$

which can also be written as

$$\sum_{S_M\in\{\pm1\}^s} \mathbb{P}(\widehat{S}_M(Y) = S_M \,|\, E_M)\, \mathbb{E}\mathcal{I}_{Y\,|\,E_M^{S_M}}(\vartheta) \preceq \mathbb{E}\,\mathcal{I}_{Y\,|\,E_M}(\vartheta).$$

In expectation, the loss of information induced by conditioning further on the vector of signs is quantified by the information $\widehat{S}_M(Y)$ carries about ϑ. Let us stress that this conclusion is only true in expectation and it may exist some vector of signs $S_M \in \{-1,+1\}^s$ such that

$$\mathcal{I}_{Y\,|\,E_M}(\vartheta) \preceq \mathcal{I}_{Y\,|\,E_M^{S_M}}(\vartheta).$$

Hence, conditioning on the signs will generally lead to wider confidence intervals. Nevertheless, let us stress that inference procedures correctly calibrated conditional on $E_M^{S_M}$ will be also valid conditional on E_M. More precisely, considering some transformation $T : \mathbb{R}^N \to \mathbb{R}$ and real valued random variables $L(Y, S_M) < U(Y, S_M)$ such that for any vector of signs $S_M \in \{-1,+1\}^s$ it holds

$$\mathbb{P}\left(T(\pi^*) \in [L(Y,S_M), U(Y,S_M)] \,|\, E_M^{S_M}\right) = 1-\alpha,$$

the confidence interval has also $(1-\alpha)$ coverage conditional on the $E_M = \{\widehat{M}(Y) = M\}$ since

$$\begin{aligned}&\mathbb{P}(T(\pi^*) \in [L(Y,\widehat{S}_M(Y)), U(Y,\widehat{S}_M(Y))] \mid E_M) \\ &= \sum_{S_M\in\{\pm1\}^s} \mathbb{P}(\widehat{S}_M(Y) = S_M \,|\, E_M)\, \underbrace{\mathbb{P}(T(\pi^*) \in [L(Y,S_M), U(Y,S_M)] \mid E_M^{S_M})}_{=1-\alpha} \\ &= 1-\alpha.\end{aligned}$$

6.9.2 Discussion in the context of the Sparse Logistic Regression

Let us recall that in Taylor and Tibshirani [2018], the authors work in the selected model for logistic regression. They consider a selected model $M \subseteq [d]$ associated to a response vector $Y = (y_i)_{i\in[n]} \in \{0,1\}^N$ where for any $i \in [N]$, y_i is a Bernoulli random variable with parameter $\{\sigma(\mathbf{X}_M\theta^*)\}_i$ for some

$\theta^* \in \mathbb{R}^s$ ($s = |M|$). As presented in Section 6.3, in Taylor and Tibshirani [2018] the authors claim the following asymptotic distribution

$$\underline{\theta} \sim \mathcal{N}(\vartheta^*_M, G_N(\vartheta^*_M)^{-1}), \tag{6.55}$$

where $\underline{\theta} = \hat{\vartheta}^\lambda_M + \lambda G_N(\widehat{\vartheta}^\lambda_M)^{-1}\widehat{S}_M(Y)$. Note that this approximation corresponds to the one usually made to form Wald tests and confidence intervals in generalized linear models. They claim that the selection event $\{Y \in \{0,1\}^N \ : \ \widehat{M}(Y) = M, \widehat{S}_M(Y) = S_M\}$ can be asymptotically approximated by

$$\{Y \ : \ \mathrm{Diag}(S_M)\left(\underline{\theta} - G_N(\vartheta^*_M)^{-1}\lambda S_M\right) \geq 0\}.$$

Let us denote by $F^{[a,b]}_{\mu,\sigma^2}$ the CDF of a $\mathcal{N}(\mu, \sigma^2)$ random variable truncated to the interval $[a, b]$. Then they use the polyhedral lemma to state that for some random variables $\mathcal{V}^-_{S_M}$ and $\mathcal{V}^+_{S_M}$ it holds

$$\left[F^{[\mathcal{V}^-_{S_M}, \mathcal{V}^+_{S_M}]}_{\vartheta^*_{M[j]}, \left[G_N(\vartheta^*_M)^{-1}\right]_{j,j}}(\underline{\theta}_j) \mid \widehat{M}(Y) = M, \ \widehat{S}_M(Y) = S_M\right] \ \sim \ \mathcal{U}([0,1]).$$

Several problems arise at this point.

1. **Lack of theoretical guarantee due to the use of Monte-Carlo estimates.**
   The first problem is that both $\underline{\theta}$ and the selection event $\{\widehat{M}(Y) = M, \ \widehat{S}_M(Y) = S_M\}$ involve the unknown parameter ϑ^*_M through $G_N(\vartheta^*_M)$. Taylor and al. propose to use a Monte-Carlo estimate for $G_N(\vartheta^*_M)$ by replacing it with $G_N(\widehat{\theta^\lambda})$. Using this Monte-Carlo estimate, one can compute L and U such that

   $$F^{[\mathcal{V}^-_{S_M}, \mathcal{V}^+_{S_M}]}_{L, \left[G_N(\vartheta^*_M)^{-1}\right]_{j,j}}(\underline{\theta}_j) = 1 - \frac{\alpha}{2} \quad \text{and} \quad F^{[\mathcal{V}^-_{S_M}, \mathcal{V}^+_{S_M}]}_{U, \left[G_N(\vartheta^*_M)^{-1}\right]_{j,j}}(\underline{\theta}_j) = \frac{\alpha}{2}.$$

   Then, $[L, U]$ is claimed to be a confidence interval with (asymptotic) $(1-\alpha)$ coverage for $\vartheta^*_{M[j]}$ conditional on $\{\widehat{M}(Y) = M, \widehat{S}_M(Y) = S_M\}$, that is,

   $$\mathbb{P}(\vartheta^*_{M[j]} \in [L, U] \mid \widehat{M}(Y) = M, \ \widehat{S}_M(Y) = S_M) = 1 - \alpha.$$

2. **Their approach is not well suited to provide more powerful inference procedures by conditioning only on E_M.**
   In the linear model, Lee et al. [2016] also start by deriving a pivotal quantity by conditioning on both the selected variables and the vector of signs. However, in the context of linear regression, the vector of signs only appears in the threshold values $\mathcal{V}^-$ and $\mathcal{V}^+$. Hence, conditioning only on the selected variables $\{\widehat{M}(Y) = M\}$ simply reduces to take the union $\cup_{S_M \in \{\pm 1\}^s}[\mathcal{V}^-_{S_M}, \mathcal{V}^+_{S_M}]$ for the truncated Gaussian. In the method proposed by Taylor and Tibshirani [2018], the vector of signs also appears in the computation of $\underline{\theta}$. The consequence is that the (asymptotic) distribution of $\underline{\theta}$ conditional on $\{\widehat{M}(Y) = M\}$ is not a truncated Gaussian anymore but a mixture of truncated Gaussians. In this situation, it seems unclear how to take advantage of this structure to provide more powerful inference procedures.

Appendix A

Markov chain theory

A.1 Introduction

We consider a state space E and a sigma-algebra Σ on E which is a standard Borel space.

Definition A.1. [Roberts and Rosenthal, 2004, section 3.2] (Markov chains)
A sequence of random variables $(X_i)_{i\geq 1}$ taking values in the measurable space (E, Σ) is a Markov chain if for any random variable $Y \in \sigma(X_i,\ i \geq n)$.

$$\mathbb{E}[Y \mid X_1, \ldots, X_n] = \mathbb{E}[Y \mid X_n].$$

Among the set of Markov chains, we will focus on the specific class of homogeneous Markov chains, namely Markov chains whose probability of transiting from one state to another does not depend on the time step n but only on the considered states. The distribution of the chain is then completely determined by its transition kernel and its initial distribution.

Definition A.2. [Meyn and Tweedie, 1993, section 3.2] (Transition kernel)
A map $P : E \times E \to [0, 1]$ is called a transition kernel if the following statements hold.

- for any $A \in \Sigma$, $P(\cdot, A)$ is a measurable map on E.
- for any $x \in E$, $P(x, \cdot)$ is a probability measure on E.

In the theory of Markov chains, two important settings need to be distinguished: the case where E is a (finite or countable) discrete space and the case where E is a continuous space. This is an important clarification since theoretical results regarding properties of Markov chains can depend of the discrete or continuous nature of the state space E. Let us also stress that in the discrete case, the transition kernel takes the form of a transition matrix and the problem of estimating P falls into the field of parametric estimation. On the contrary when E is continuous, non-parametric methods are required. In this thesis, we tackle the two different settings. In Chapter 2 we work with Markov chains on a finite and discrete state space, in Chapter 3 we consider general state space (discrete or continuous), while our results from Chapters 4 hold for general state spaces.

Definition A.3. The random process $(X_i)_{i\geq 1}$ is a homogenous Markov chain with initial distribution χ and transition kernel P if for any $n \in \mathbb{N}^*$ and for any $A_1, \ldots, A_n \in \Sigma$,

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{x_1 \in A_1} \cdots \int_{x_{n-1} \in A_{n-1}} \chi(dx_1) P(x_1, dx_2) \ldots P(x_{n-1}, A_n).$$

In the following, we denote by $(X_i)_{i\geq 1}$ a time homogeneous Markov chain on the state space (E, Σ) with transition kernel P.

Definition A.4. A probability measure π on (E, Σ) is said to be a stationary measure for the Markov chain $(X_i)_{i\geq 1}$ if

$$\forall A \in \Sigma, \quad \int_E \pi(dx) P(x, A) = \pi(A).$$

A.2 Ergodic and reversible Markov chains

Definition A.5. [Roberts and Rosenthal, 2004, section 3.2] (ϕ-irreducible Markov chains)
The Markov chain $(X_i)_{i\geq 1}$ is said ϕ-irreducible if there exists a non-zero σ-finite measure ϕ on E such that for all $A \in \Sigma$ with $\phi(A) > 0$, and for all $x \in E$, there exists a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$ (where $P^n(x, \cdot)$ denotes the distribution of X_{n+1} conditioned on $X_1 = x$).

Definition A.6. [Roberts and Rosenthal, 2004, section 3.2] (Aperiodic Markov chains)
The Markov chain $(X_i)_{i\geq 1}$ with stationary distribution π is aperiodic if there do not exist $m \geq 2$ and disjoint subsets $A_1, \dots, A_m \subset E$ with $P(x, A_{i+1}) = 1$ for all $x \in A_i$ $(1 \leq i \leq m-1)$, and $P(x, A_1) = 1$ for all $x \in A_m$, such that $\pi(A_1) > 0$ (and hence $\pi(A_i) > 0$ for all i).

Definition A.7. [Roberts and Rosenthal, 2004, section 3.4] (Geometric ergodicity)
The Markov chain $(X_i)_{i\geq 1}$ is said geometrically ergodic if there exists a stationary distribution π, $\rho \in (0, 1)$ and $C : E \to [1, \infty)$ such that

$$\|P^n(x, \cdot) - \pi\|_{\mathrm{TV}} \leq C(x)\rho^n, \qquad \forall n \geq 0,\ \pi\text{−a.e } x \in E,$$

where $\|\mu\|_{\mathrm{TV}} := \sup_{A\in\Sigma} |\mu(A)|$.

Definition A.8. [Roberts and Rosenthal, 2004, section 3.3] and [Meyn and Tweedie, 1993, Chapter 16] (Uniform ergodicity)
The Markov chain $(X_i)_{i\geq 1}$ is said uniformly ergodic if there exists an stationary distribution π and constants $0 < \rho < 1$ and $L > 0$ such that

$$\|P^n(x, \cdot) - \pi\|_{\mathrm{TV}} \leq L\rho^n, \qquad \forall n \geq 0,\ \pi\text{−a.e } x \in E,$$

where $\|\mu\|_{\mathrm{TV}} := \sup_{A\in\Sigma} |\mu(A)|$.
Equivalently, the Markov chain $(X_i)_{i\geq 1}$ is uniformly ergodic if the whole space E is a small set, namely if there exist an integer $m \geq 1$, $\delta_m > 0$ and a probability measure ν such that

$$\forall x \in E,\ \forall A \in \Sigma, \quad P^m(x, A) \geq \delta_m \nu(A).$$

Remark. A Markov chain geometrically or uniformly ergodic admits a unique stationary distribution and is aperiodic.

Definition A.9. (Reversible Markov chain)
A Markov chain is said reversible if there exists a distribution π satisfying

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx). \tag{A.1}$$

A straightforward remark is that any probability measure π that satisfies Eq.(A.1) is a stationary measure of the chain.

A.3 Spectral gap

This section is largely inspired from Fan et al. [2021] and [Jiang et al., 2018, Section 2.1]. Let us consider that the Markov chain $(X_i)_{i\geq 1}$ admits a unique stationary distribution π on E. For any real-valued, Σ-measurable function $h : E \to \mathbb{R}$, we define $\pi(h) := \int h(x)\pi(dx)$. The set

$$L^2(E, \Sigma, \pi) := \{h : \pi(h^2) < \infty\}$$

is a Hilbert space endowed with the inner product

$$\langle h_1, h_2\rangle_\pi = \int h_1(x)h_2(x)\pi(dx),\ \forall h_1, h_2 \in L^2(E, \Sigma, \pi).$$

The map

$$\|\cdot\|_\pi : h \in L^2(E, \Sigma, \pi) \mapsto \|h\|_\pi = \sqrt{\langle h, h\rangle_\pi},$$

is a norm on $L^2(E,\Sigma,\pi)$. $\|\cdot\|_\pi$ naturally allows to define the norm of a linear operator T on $L^2(E,\Sigma,\pi)$ as

$$N_\pi(T) = \sup\{\|Th\|_\pi : \|h\|_\pi = 1\}.$$

To each transition probability kernel $P(x,B)$ with $x \in E$ and $B \in \Sigma$ stationary with respect to π, we can associate a bounded linear operator $h \mapsto \int h(y)P(\cdot,dy)$ on $L^2(E,\Sigma,\pi)$. Denoting this operator P, we get

$$Ph(x) = \int h(y)P(x,dy),\ \forall x \in E,\ \forall h \in L^2(E,\Sigma,\pi).$$

Let $L_0^2(\pi) := \{h \in L^2(E,\Sigma,\pi)\ :\ \pi(h) = 0\}$. We define the absolute spectral gap of a Markov operator.

Definition A.10. (Absolute spectral gap) The Markov operator P admits an absolute spectral gap $1-\lambda$ if

$$\lambda := \sup\left\{\frac{\|Ph\|_\pi}{\|h\|_\pi}\ :\ h \in L_0^2(\pi),\ h \neq 0\right\} < 1.$$

The next result provides a connection between the existence of an absolute spectral gap and uniform ergodicity.

Proposition A.11. *[Ferré et al., 2012, section 2.3]*
A uniformly ergodic Markov chain admits an absolute spectral gap.

Denoting by P^* the adjoint or time-reversal operator of the Markov operator P, we can define the self-adjoint operator $R = (P+P^*)/2$. The spectrum of a self-adjoint Markov operator like R acting on $L_0^2(\pi)$ is contained in $[-1,+1]$. The gap between 1 and the maximum of the spectrum of R is called the right L^2-spectral gap of P.

Definition A.12. (Right L^2-spectral gap) The Markov operator P has right L^2-spectral gap $1-\lambda_+(R)$ if the operator $R = (P+P^*)/2$ satisfies

$$\lambda_+(R) := \sup\{s : s \in \text{spectrum of } R \text{ acting on } L_0^2(\pi)\} < 1.$$

Note that it holds (cf. [Jiang et al., 2018, Section 2]),

$$0 \leq \lambda \leq 1, \quad -1 \leq \lambda_+(R) \leq \lambda \quad \text{and} \quad \max(\lambda_+(R),0) \leq \lambda.$$

A.4 Splitting technique

We assume that the Markov chain $(X_i)_{i\geq 1}$ is uniformly ergodic (see Definition A.8). We extend the Markov chain $(X_i)_{i\geq 1}$ to a new (so called *split*) chain $(\widetilde{X}_n, R_n) \in E \times \{0,1\}$ (see [Meyn and Tweedie, 1993, Section 5.1] for a reminder on the splitting technique), satisfying the following properties.

- $(\widetilde{X}_n)_n$ is again a Markov chain with transition kernel P with the same initial distribution as $(X_n)_n$. We recall that π is the stationary distribution on the E.
- if we define $T_1 = \inf\{n > 0 : R_{nm} = 1\}$,

  $$T_{i+1} = \inf\{n > 0 : R_{(T_1+\cdots+T_i+n)m} = 1\},$$

  then $T_1, T_2, \ldots$ are well defined and independent. Moreover $T_2, T_3, \ldots$ are i.i.d.
- if we define $S_i = T_1 + \cdots + T_i$, then the "blocks"

  $$Y_0 = (\widetilde{X}_1, \ldots, \widetilde{X}_{mT_1+m-1}), \quad \text{and} \quad Y_i = (\widetilde{X}_{m(S_i+1)}, \ldots, \widetilde{X}_{m(S_{i+1}+1)-1}),\ i > 0,$$

  form a one-dependent sequence (i.e. for all i, $\sigma((Y_j)_{j<i})$ and $\sigma((Y_j)_{j>i})$ are independent). Moreover, the sequence $Y_1, Y_2, \ldots$ is stationary and if $m = 1$ the variables $Y_0, Y_1, \ldots$ are independent. In consequence, for any measurable space $(S,\mathcal{B})$ and measurable functions $f : S \to \mathbb{R}$, the variables

  $$Z_i = Z_i(f) = \sum_{j=m(S_i+1)}^{m(S_{i+1}+1)-1} f(\widetilde{X}_j), \quad i \geq 1,$$

constitute a one-dependent sequence (an i.i.d. sequence if $m = 1$). Additionally, if f is π-integrable (recall that π is the unique stationary measure for the chain), then

$$\mathbb{E}[Z_i] = \delta_m^{-1} m \int f d\pi.$$

- the distribution of T_1 depends only on π, P, δ_m, μ, whereas the law of T_2 only on P, δ_m and μ.

Remarks.

- Let us highlight that $(\widetilde{X}_n)_n$ is a Markov chain with transition kernel P and same initial distribution as $(X_n)_n$.
- Let us mention that we prove in Chapter 4 (cf. Proposition 4.2) that for any uniformly ergodic Markov chain $(X_i)_{i\geq 1}$, the regeneration times are exponentially integrable. This means that for some $\tau \in (0, \infty)$

$$\|T_1\|_{\psi_1} < \tau \quad \text{and} \quad \|T_2\|_{\psi_1} < \tau,$$

where $\|\cdot\|_{\psi_1}$ is the 1-Orlicz norm presented in Definition A.13.

Definition A.13. For $\alpha > 0$, define the function $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ with the formula $\psi_\alpha(x) = \exp(x\alpha) - 1$. Then for a random variable X, the α-Orlicz norm is given by

$$\|X\|_{\psi_\alpha} = \inf \{\gamma > 0 : \mathbb{E}[\psi_\alpha(|X|/\gamma)] \leq 1\}.$$

A.5 Concentration lemmas for Markov chains

A.5.1 Hoeffding inequality for uniformly ergodic Markov chains

Proposition A.14 is an Hoeffding bound for uniformly ergodic Markov chains. Some Hoeffding inequalities for uniformly ergodic Markov chains without condition on the initial distribution already exist (see Glynn and Ormoneit [2002] or Boucher [2009]), but they require n to be large enough to hold or can involve quantities related to the chain that we do not use in this thesis (such that the Drazin inverse of $I - P$). This is the reason why we propose here to present a different Hoeffding inequality for uniformly ergodic Markov chains that holds for any sample size n, any initial distribution of the chain and that only uses the notations of Section 4.2 from Chapter 4.

Proposition A.14. *Let $(X_i)_{i\geq 1}$ be a Markov chain on E uniformly ergodic (namely satisfying Assumption 1 from Chapter 4) with stationary distribution π and let us consider some function $f : E \to \mathbb{R}$ such that $\mathbb{E}_{X\sim\pi}[f(X)] = 0$. Then it holds for any $t \geq 0$*

$$\mathbb{P}\left(\left|\sum_{i=1}^n f(X_i)\right| \geq t\right) \leq 16 \exp\left(-\frac{1}{K(m,\tau)} \frac{t^2}{n\|f\|_\infty^2}\right),$$

where $K(m, \tau) = 2Km^2\tau^2$ for some universal constant $K > 0$. We refer to Assumption 1 and the following remark (or to [Duchemin et al., 2022b, Section 2]) for the definitions of the constants m and τ.

Proof of Proposition A.14. Let us first recall that under Assumption 1, the 1-Orlicz norm of the regeneration times of the split chain are bounded by some finite constant $\tau > 0$ (see the remark at the end of Section A.4 and Definition A.14). In this proof, we will use the notations introduced in Section A.4. Since the chain $(\widetilde{X}_n)_n$ is distributed as $(X_i)_{i\geq 1}$, we will identify $(\widetilde{X}_i)_{i\geq 1}$ and $(X_i)_{i\geq 1}$ in the proof.
Let us consider $N = \sup\{i \in \mathbb{N} \ : \ mS_{i+1} + m - 1 \leq n\}$. Then,

$$\Big|\sum_{i=1}^n f(X_i)\Big| = \Big|\sum_{l=0}^N Z_l + \sum_{i=m(S_N+1)}^n f(X_i)\Big| \leq \Big|\sum_{l=0}^{\lfloor N/2\rfloor} Z_{2l}\Big| + \Big|\sum_{l=0}^{\lfloor (N-1)/2\rfloor} Z_{2l+1}\Big| + \Big|\sum_{i=m(S_N+1)}^n f(X_i)\Big|. \quad \text{(A.2)}$$

We have $|\sum_{i=m(S_N+1)}^n f(X_i)| \leq AmT_{N+1}$. So using the definition of the Orlicz norm and the fact that

the random variables $(T_i)_{i\geq 2}$ are i.i.d., it holds for any $t \geq 0$,

$$\mathbb{P}\big(\big|\sum_{i=m(S_N+1)}^{n} f(X_i)\big| \geq t\big) \leq \mathbb{P}(T_{N+1} \geq \frac{t}{Am}) \leq \mathbb{P}(\max(T_1, T_2) \geq \frac{t}{Am}) \leq 4\exp\big(-\frac{t}{Am\tau}\big).$$

In order to control the first two terms in (A.2), we need to describe the tail behaviour of the random variable N with Lemma A.15.

Lemma A.15. *(cf. [Adamczak, 2008, Lemma 5])*
We denote $R = \lfloor 3n/(\mathbb{E}T_2)\rfloor$. If $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} \leq \tau$, then $\mathbb{P}(N > R) \leq 2\exp\big(-\frac{n\mathbb{E}T_2}{8\tau^2}\big)$.

The random variable Z_{2l} is $\sigma(X_{m(S_{2l}+1)}, \dots, X_{m(S_{2l+1}+1)-1})$-measurable. Hence the random variables $(Z_{2l})_l$ are independent (see Section 4.2.3). Moreover, one has that for any l, $\mathbb{E}[Z_{2l}] = 0$. This is due to [Meyn and Tweedie, 1993, Eq.(17.23) Theorem 17.3.1] together with the assumption that $\mathbb{E}_{X\sim\pi}[f(X)] = 0$. Let us finally notice for any $l \geq 0$, $|Z_{2l}| \leq AmT_{2l+1}$, so $\|Z_{2l}\|_{\psi_1} \leq Am\max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1}) \leq Am\tau$. One can similarly get that $(Z_{2l+1})_l$ are independent with $\mathbb{E}[Z_{2l+1}] = 0$ and $\|Z_{2l+1}\|_{\psi_1} \leq Am\tau$ for all $l \in \mathbb{N}$. Using these facts we have for any $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}\big(\big|\sum_{l=0}^{\lfloor N/2\rfloor} Z_{2l}\big| + \big|\sum_{l=0}^{\lfloor (N-1)/2\rfloor} Z_{2l+1}\big| \geq t\big)\\
&\leq \mathbb{P}\big(\big|\sum_{l=0}^{\lfloor N/2\rfloor} Z_{2l}\big| + \big|\sum_{l=0}^{\lfloor (N-1)/2\rfloor} Z_{2l+1}\big| \geq t\,, N \leq R\big) + 2\exp\big(-\frac{n\mathbb{E}T_2}{8\tau^2}\big)\\
&\leq \mathbb{P}\big(\max_{0\leq s\leq \lfloor R/2\rfloor}\big|\sum_{l=0}^{s} Z_{2l}\big| \geq t/2\big) + \mathbb{P}\big(\max_{0\leq s\leq \lfloor (R-1)/2\rfloor}\big|\sum_{l=0}^{s} Z_{2l+1}\big| \geq t/2\big) + 2\exp\big(-\frac{n\mathbb{E}T_2}{8\tau^2}\big)\\
&\leq 3\mathbb{P}\big(\big|\sum_{l=0}^{\lfloor R/2\rfloor} Z_{2l}\big| \geq t/6\big) + 3\mathbb{P}\big(\big|\sum_{l=0}^{\lfloor (R-1)/2\rfloor} Z_{2l+1}\big| \geq t/6\big) + 2\exp\big(-\frac{n\mathbb{E}T_2}{8\tau^2}\big) \quad \text{(Using Lemma A.16)}\\
&\leq 12\exp\big(-\frac{1}{8}\min\big(\frac{t^2}{36RA^2m^2\tau^2}, \frac{t}{6Am\tau}\big)\big) + 2\exp\big(-\frac{n\mathbb{E}T_2}{8\tau^2}\big),
\end{aligned}$$

where we used Lemma 4.13 in the last inequality.

Lemma A.16. *(cf. [Kwapień and Woyczyński, 1992, Proposition 1.1.1]) If $X_1, X_2, \dots$ are independent Banach space valued random variables (not necessarily identically distributed), and if $S_k = \sum_{i=1}^{k} X_i$, then*

$$\mathbb{P}\left(\max_{1\leq j\leq k}\|S_j\| > t\right) \leq 3\max_{1\leq j\leq k}\mathbb{P}\left(\|S_j\| > t/3\right).$$

Gathering the previous results, we obtain that for any $t \geq 0$

$$\begin{aligned}
\mathbb{P}\big(\big|\sum_{i=1}^{n} f(X_i)\big| \geq t\big) \leq\ & 12\exp\big(-\frac{1}{8}\min\big(\frac{t^2(\mathbb{E}T_2)}{36\times 12\times nA^2m^2\tau^2}, \frac{t}{12Am\tau}\big)\big)\\
&+ 2\exp\big(-\frac{n\mathbb{E}T_2}{8\tau^2}\big) + 4\exp\big(-\frac{t}{2Am\tau}\big).
\end{aligned}$$

Since the left hand side of the previous inequality is zero for $t \geq nA$, and since $m \geq 1$, we obtain Proposition A.14. □

A.5.2 Bernstein's inequality for non-stationary Markov chains

Proposition A.17 is an extension of the Bernstein type concentration inequality from Jiang et al. [2018] to non-stationary Markov chains. We provide a proof of this result in this section.

Proposition A.17. *Suppose that the sequence $(X_i)_{i\geq 1}$ is a Markov chain satisfying Assumptions 1 and 4 from Chapter 4 with stationary distribution π and with an absolute spectral gap $1-\lambda > 0$ (cf. Definition A.10). Let us consider some $n \in \mathbb{N}\backslash\{0\}$ and bounded real valued functions $(f_i)_{1\leq i\leq n}$ such that for*

any $i \in \{1, \dots, n\}$, $\int f_i(x)d\pi(x) = 0$ *and* $\|f_i\|_\infty \leq c$ *for some* $c > 0$. *Let* $\sigma^2 = \sum_{i=1}^n \int f_i^2(x)d\pi(x)/n$. *Then for any* $\epsilon \geq 0$ *it holds*

$$\mathbb{P}\left(\sum_{i=1}^n f_i(X_i) \geq \epsilon\right) \leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} \exp\left(-\frac{\epsilon^2/(2q)}{A_2\sigma^2 + A_1 c\epsilon}\right),$$

where $A_2 := \frac{1+\lambda}{1-\lambda}$ *and* $A_1 := \frac{1}{3}\mathbb{1}_{\lambda=0} + \frac{5}{1-\lambda}\mathbb{1}_{\lambda>0}$. q *is the constant introduced in Assumption 4. Stated otherwise, for any* $u > 0$ *it holds*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f_i(X_i) > \frac{2quA_1c}{n} + \sqrt{\frac{2quA_2\sigma^2}{n}}\right) \leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} e^{-u}.$$

Proof of Proposition A.17. Let us recall that we denote indifferently $\mathbb{P}_\chi$ or $\mathbb{P}$ the probability distribution of the Markov chain $(X_i)_{i\geq 1}$ when the distribution of the first state X_1 is χ, whereas $\mathbb{P}_\pi$ refers to the distribution of the Markov chain when the distribution of the first state X_1 is the invariant measure π. In Jiang et al. [2018], they proved that for any $0 \leq t < (1-\lambda)/(5c)$, it holds

$$\mathbb{E}_\pi\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] \leq \exp\left(\frac{n\sigma^2}{c^2}(e^{tc} - tc - 1) + \frac{n\sigma^2\lambda t^2}{1-\lambda-5ct}\right).$$

We deduce that for any $0 \leq t < (1-\lambda)/(5cq)$,

$$\begin{aligned}
\mathbb{E}_\chi\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] &\leq \mathbb{E}_\pi\left[\frac{d\chi}{d\pi}(X_1)e^{t\sum_{i=1}^n f_i(X_i)}\right] \\
&\leq \left\{\mathbb{E}_\pi\left[\left|\frac{d\chi}{d\pi}(X_1)\right|^p\right]\right\}^{1/p}\left\{\mathbb{E}_\pi\left[e^{qt\sum_{i=1}^n f_i(X_i)}\right]\right\}^{1/q} \\
&= \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\left\{\mathbb{E}_\pi\left[e^{qt\sum_{i=1}^n f_i(X_i)}\right]\right\}^{1/q} \\
&\leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\left\{\exp\left(\frac{n\sigma^2}{c^2}(e^{tqc} - tqc - 1) + \frac{n\sigma^2\lambda q^2t^2}{1-\lambda-5cqt}\right)\right\}^{1/q} \\
&= \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\exp\left(\frac{n\sigma^2}{qc^2}(e^{tqc} - tqc - 1) + \frac{n\sigma^2\lambda qt^2}{1-\lambda-5cqt}\right). \qquad \text{(A.3)}
\end{aligned}$$

Let us define

$$g_1(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{n\sigma^2}{qc^2}(e^{tqc} - tqc - 1) & \text{if } t \geq 0 \end{cases}$$

and

$$g_2(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{n\sigma^2\lambda qt^2}{1-\lambda-5cqt} & \text{if } 0 \leq t < \frac{1-\lambda}{5cq} \\ +\infty & \text{if } t \geq \frac{1-\lambda}{5cq} \end{cases}.$$

In order to lower-bound the convex conjugate of the function $g_1 + g_2$, we will need the convex conjugate of g_1 and g_2 which are provided by Lemma A.18. The proof of Lemma A.18 can be found in Section A.5.2.1.

Lemma A.18. *g_1 and g_2 are closed proper convex functions with convex conjugates*

$$\forall \epsilon_1 \in \mathbb{R}, \quad g_1^*(\epsilon_1) = \begin{cases} \frac{n\sigma^2}{qc^2}h_1\left(\frac{\epsilon_1 c}{n\sigma^2}\right) & \text{if } \epsilon_1 \geq 0 \\ +\infty & \text{if } \epsilon_1 < 0 \end{cases} \qquad \text{(A.4)}$$

with $h_1(u) = (1+u)\log(1+u) - u \geq \frac{u^2}{2(1+u/3)}$ *for any* $u \geq 0$, *and*

$$\forall \epsilon_2 \in \mathbb{R}, \quad g_2^*(\epsilon_2) = \begin{cases} \frac{(1-\lambda)\epsilon^2}{qn\sigma^2\lambda} h_2(\frac{5c\epsilon_2}{n\sigma^2\lambda}) & \text{if } \epsilon_2 \geq 0 \\ +\infty & \text{if } \epsilon_2 < 0 \end{cases} \tag{A.5}$$

with $h_2(u) = \left(\frac{(\sqrt{u+1}-1}{u}\right)^2 \geq \frac{1}{2(u+2)}$.

Since $g_1(t) = O(t^2)$ and $g_2(t) = O(t^2)$ as $t \to 0^+$, $t\epsilon - g_1(t) - g_2(t) > 0$ for small enough $t > 0$, and $t\epsilon - g_1(t) - g_2(t) \leq 0$ for $t \leq 0$. Hence

$$(g_1 + g_2)^*(\epsilon) = \sup_{0 \leq t < (1-\lambda)/(5cq)} \epsilon t - g_1(t) - g_2(t) = \sup_{t \in \mathbb{R}} \epsilon t - g_1(t) - g_2(t).$$

• If $\lambda > 0$, then by the Moreau-Rockafellar formula [Rockafellar, 1970, Theorem 16.4], the convex conjugate of $g_1 + g_2$ is the infimal convolution of their conjugates g_1^* and g_2^*, namely

$$(g_1 + g_2)^*(\epsilon) = \inf \left\{ g_1^*(\epsilon_1) + g_2^*(\epsilon_2) : \epsilon = \epsilon_1 + \epsilon_2,\ \epsilon_1, \epsilon_2 \in \mathbb{R} \right\}.$$

Using (A.4) and (A.5), this reads as

$$(g_1 + g_2)^*(\epsilon) = \inf \left\{ \frac{n\sigma^2}{qc^2} h_1\left(\frac{\epsilon_1 c}{n\sigma^2}\right) + \frac{(1-\lambda)\epsilon_2^2}{qn\sigma^2\lambda} h_2\left(\frac{5c\epsilon_2}{n\sigma^2\lambda}\right) : \epsilon = \epsilon_1 + \epsilon_2,\ \epsilon_1, \epsilon_2 \geq 0 \right\}.$$

Bounding $h_1(u) \geq \frac{u^2}{2(1+u/3)}$ and $h_2(u) \geq \frac{1}{2(u+2)}$, we have

$$\begin{aligned}(g_1 + g_2)^*(\epsilon) &\geq \inf \left\{ \frac{1}{qc^2} \frac{c^2\epsilon_1^2}{2(n\sigma^2 + c\epsilon_1/3)} + \frac{(1-\lambda)\epsilon_2^2}{qn\sigma^2\lambda} \frac{1}{\frac{10c\epsilon_2}{n\sigma^2\lambda} + 4} : \epsilon = \epsilon_1 + \epsilon_2,\ \epsilon_1, \epsilon_2 \geq 0 \right\} \\ &\geq \inf \left\{ \frac{\epsilon_1^2}{2q(n\sigma^2 + c\epsilon_1/3)} + \frac{(1-\lambda)\epsilon_2^2}{2q} \frac{1}{5c\epsilon_2 + 2n\sigma^2\lambda} : \epsilon = \epsilon_1 + \epsilon_2,\ \epsilon_1, \epsilon_2 \geq 0 \right\}.\end{aligned}$$

Using the fact that $\epsilon_1^2/a + \epsilon_2^2/b \geq (\epsilon_1 + \epsilon_2)^2/(a+b)$ for any non-negative ϵ_1, ϵ_2 and positive a, b yield

$$\begin{aligned}(g_1 + g_2)^*(\epsilon) &\geq \inf \left\{ \frac{(\epsilon_1 + \epsilon_2)^2}{2q(n\sigma^2 + c\epsilon_1/3) + \frac{2q}{(1-\lambda)}(5c\epsilon_2 + 2n\sigma^2\lambda)} : \epsilon = \epsilon_1 + \epsilon_2,\ \epsilon_1, \epsilon_2 \geq 0 \right\} \\ &= \inf \left\{ \frac{\epsilon^2/(2q)}{\frac{1+\lambda}{1-\lambda} n\sigma^2 + c\epsilon_1/3 + \frac{5c\epsilon}{1-\lambda} - \frac{5c\epsilon_1}{1-\lambda}} : \epsilon = \epsilon_1 + \epsilon_2,\ \epsilon_1, \epsilon_2 \geq 0 \right\} \\ &\geq \frac{\epsilon^2/(2q)}{\frac{1+\lambda}{1-\lambda} n\sigma^2 + \frac{5c\epsilon}{1-\lambda}},\end{aligned}$$

where we used for the last inequality that for any $\epsilon_1 \geq 0$,

$$c\epsilon_1/3 - \frac{5c\epsilon_1}{1-\lambda} = \frac{c\epsilon_1}{3(1-\lambda)}(1 - \lambda - 15) < 0.$$

• If $\lambda = 0$,

$$(g_1 + g_2)^*(\epsilon) = g_1^*(\epsilon) = \frac{n\sigma^2}{qc^2} h_1\left(\frac{\epsilon_1 c}{n\sigma^2}\right) \geq \frac{\epsilon^2/(2q)}{n\sigma^2 + c\epsilon/3}.$$

We deduce from the previous computations that for any $t, \epsilon \geq 0$ it holds

$$
\begin{aligned}
&\mathbb{P}_\chi\left(\sum_{i=1}^n f_i(X_i) \geq \epsilon\right) \\
&\leq e^{-\epsilon t}\mathbb{E}_\chi\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] \text{ using Markov's inequality} \\
&\leq e^{-\epsilon t}\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\exp\left(\frac{n\sigma^2}{qc^2}(e^{tqc}-tqc-1)+\frac{n\sigma^2\lambda qt^2}{1-\lambda-5cqt}\right) \text{ using (A.3)} \\
&\leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\exp\left(-(g_1+g_2)^*(\epsilon)\right) \\
&\leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\times\begin{cases}\exp\left(-\frac{\epsilon^2/(2q)}{\frac{1+\lambda}{1-\lambda}n\sigma^2+\frac{5c\epsilon}{1-\lambda}}\right) & \text{if } \lambda>0\\ \exp\left(-\frac{\epsilon^2/(2q)}{n\sigma^2+c\epsilon/3}\right) & \text{if } \lambda=0\end{cases}.
\end{aligned}
$$

A.5.2.1 Proof of Lemma A.18

The convex conjugate of g_1 is usual and follows from easy computations. We focus on the convex conjugate of g_2 which requires non-trivial computations.
Let $f_\epsilon(t)=\epsilon t-\frac{n\sigma^2\lambda qt^2}{1-\lambda-5cqt}$ for any $0\leq t<(1-\lambda)/(5cq)$. We have for any $0\leq t<(1-\lambda)/(5cq)$,

$$
f_\epsilon'(t)=\epsilon-\frac{2n\sigma^2\lambda qt(1-\lambda-5cqt)+5cq^2n\sigma^2\lambda t^2}{(1-\lambda-5cqt)^2}.
$$

Hence, for $0\leq t<(1-\lambda)/(5cq)$ such that $f_\epsilon'(t)=0$ we have

$$
\begin{aligned}
&\epsilon(1-\lambda-5cqt)^2-2n\sigma^2\lambda qt(1-\lambda-5cqt)-5cq^2n\sigma^2\lambda t^2 &&=0\\
\Leftrightarrow\quad &\epsilon(1-\lambda)^2-10\epsilon(1-\lambda)cqt+25\epsilon c^2q^2t^2-2n\sigma^2\lambda q(1-\lambda)t+10n\sigma^2cq^2\lambda t^2-5n\sigma^2cq^2\lambda t^2 &&=0\\
\Leftrightarrow\quad &\epsilon(1-\lambda)^2-10\epsilon(1-\lambda)cqt+25\epsilon c^2q^2t^2-2n\sigma^2\lambda q(1-\lambda)t+5n\sigma^2cq^2\lambda t^2 &&=0.
\end{aligned}
$$

We are looking for the roots of a polynomial of degree 2 in t. The discriminant is

$$
\begin{aligned}
\Delta&=(10\epsilon(1-\lambda)cq+2n\sigma^2\lambda q(1-\lambda))^2-4\epsilon(1-\lambda)^2(25\epsilon c^2q^2+5n\sigma^2cq^2\lambda)\\
&=4(1-\lambda)^2q^2\left[(5\epsilon c+n\sigma^2\lambda)^2-\epsilon(25c^2\epsilon+5n\sigma^2c\lambda)\right]\\
&=4(1-\lambda)^2q^2\left[25\epsilon^2c^2+10n\sigma^2\lambda\epsilon c+n^2\sigma^4\lambda^2-25c^2\epsilon^2-5n\sigma^2c\lambda\epsilon\right]\\
&=4(1-\lambda)^2q^2\left[5n\sigma^2\lambda\epsilon c+n^2\sigma^4\lambda^2\right]\\
&=4(1-\lambda)^2q^2n^2\sigma^4\lambda^2\left[u+1\right],
\end{aligned}
$$

where $u=\frac{5c\epsilon}{n\sigma^2\lambda}$.
Hence, the roots of the polynomial of interest are of the form

$$
\begin{aligned}
&\frac{10\epsilon(1-\lambda)cq+2n\sigma^2\lambda q(1-\lambda)\pm\sqrt{\Delta}}{2\left[25\epsilon c^2q^2+5n\sigma^2cq^2\lambda\right]}\\
&=\frac{2(1-\lambda)qn\sigma^2\lambda\left[\frac{5\epsilon c}{n\sigma^2\lambda}+1\right]\pm\sqrt{\Delta}}{10q^2cn\sigma^2\lambda\left[\frac{5\epsilon c}{n\sigma^2\lambda}+1\right]}\\
&=\frac{1-\lambda}{5cq}\times\frac{u+1\pm\sqrt{u+1}}{u+1}.
\end{aligned}
$$

We deduce that the polynomial has a root in the interval $[0,\frac{1-\lambda}{5cq})$ which is given by

$$
t^*:=\frac{1-\lambda}{5cq}\times\frac{u+1-\sqrt{u+1}}{u+1},
$$

and one can check that this critical point corresponds to a maximum of the function f_ϵ. We deduce that

for any $\epsilon > 0$,

$$\begin{aligned}
g_2^*(\epsilon) &= \epsilon t^* - \frac{n\sigma^2\lambda q (t^*)^2}{1-\lambda-5cqt^*} \\
&= t^* \left\{ \epsilon - \frac{n\sigma^2\lambda q \frac{1-\lambda}{5cq} \times \frac{u+1-\sqrt{u+1}}{u+1}}{1-\lambda-5cq\frac{1-\lambda}{5cq} \times \frac{u+1-\sqrt{u+1}}{u+1}} \right\} = t^* \left\{ \epsilon - \frac{n\sigma^2\lambda q \left(u+1-\sqrt{u+1}\right)}{5cq(u+1) - 5cq\left(u+1-\sqrt{u+1}\right)} \right\} \\
&= t^* \left\{ \epsilon - \frac{n\sigma^2\lambda q \left(u+1-\sqrt{u+1}\right)}{5cq\sqrt{u+1}} \right\} = t^* \left\{ \epsilon - \frac{n\sigma^2\lambda}{5c} \times \frac{\left(u+1-\sqrt{u+1}\right)}{\sqrt{u+1}} \right\} \\
&= t^* \left\{ \epsilon - \frac{\epsilon\left(u+1-\sqrt{u+1}\right)}{u\sqrt{u+1}} \right\} = t^*\epsilon \left\{ \frac{u\sqrt{u+1}-u-1+\sqrt{u+1}}{u\sqrt{u+1}} \right\} \\
&= \frac{1-\lambda}{5cq}\epsilon \times \frac{u+1-\sqrt{u+1}}{u+1} \left\{ \frac{u-\sqrt{u+1}+1}{u} \right\} \\
&= \frac{(1-\lambda)\epsilon}{q}\frac{1}{5c} \times \left(\sqrt{u+1}-1\right) \left\{ \frac{\sqrt{u+1}-1}{u} \right\} = \frac{(1-\lambda)\epsilon^2}{qn\sigma^2\lambda}\frac{(\sqrt{u+1}-1)^2}{u^2} = \frac{(1-\lambda)\epsilon^2}{qn\sigma^2\lambda}h_2(u),
\end{aligned}$$

where $h_2(u) = \frac{(\sqrt{u+1}-1)^2}{u^2}$. However, the function $u \mapsto \sqrt{u+1}$ is analytic on $]0,+\infty[$ and for any $v \in]0,+\infty[$,

$$\sqrt{1+v} = \sum_{k=0}^{\infty} \frac{v^k}{k!} a_k,$$

where $a_0 = 1$ and for all $k \in \mathbb{N}^*$, $a_k = \frac{1}{2}(\frac{1}{2}-1)\dots(\frac{1}{2}-k+1)$. Hence, we have

$$\begin{aligned}
\frac{\sqrt{v+1}-1}{v} &= \sum_{k=1}^{\infty} \frac{v^{k-1}}{k!}\frac{1}{2}(\frac{1}{2}-1)\dots(\frac{1}{2}-k+1) = \sum_{k=0}^{\infty} \frac{v^k}{(k+1)!}\frac{1}{2}(\frac{1}{2}-1)\dots(\frac{1}{2}-k) \\
&= \frac{1}{2}\sum_{k=0}^{\infty} \frac{v^k}{(k+1)!} b_k = \frac{1}{2}\sum_{k=0}^{\infty} \frac{(v/2)^k}{k!} b_k \underbrace{\frac{2^k}{k+1}}_{\geq 1} \\
&\geq \frac{1}{2}\sum_{k=0}^{\infty} \frac{(v/2)^k}{k!} b_k = \frac{1}{2}(v/2+1)^{-1/2} = \frac{1}{\sqrt{2}}\frac{1}{\sqrt{v+2}},
\end{aligned}$$

where we have denoted $b_0 = 1$ and for all $k \in \mathbb{N}^*$, $b_k = \left(-\frac{1}{2}\right)\left(-\frac{1}{2}-1\right)\dots\left(-\frac{1}{2}-k+1\right)$. Hence we proved that for any $\epsilon > 0$,

$$g_2^*(\epsilon) = \frac{(1-\lambda)\epsilon^2}{qn\sigma^2\lambda}h_2(u) \geq \frac{(1-\lambda)\epsilon^2}{2qn\sigma^2\lambda} \times \frac{1}{u+2} \text{ with } u = \frac{5c\epsilon}{n\sigma^2\lambda}.$$

□

Bibliography

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe and Colin Sandon. Community detection in general Stochastic Block Models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015a.

Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015b.

Emmanuel Abbe and Colin Sandon. Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/6c29793a140a811d0c45ce03c1c93a28-Paper.pdf.

Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13(34):1000–1034, 2008. doi: 10.1214/EJP.v13-521. URL https://doi.org/10.1214/EJP.v13-521.

Radosław Adamczak and Witold Bednorz. Some remarks on MCMC estimation of spectra of integral operators. *Bernoulli*, 21(4):2073—-2092, Nov 2015a. ISSN 1350-7265. doi: 10.3150/14-bej635. URL http://dx.doi.org/10.3150/14-BEJ635.

Radosław Adamczak and Witold Bednorz. Exponential concentration inequalities for additive functionals of Markov chains. *ESAIM: Probability and Statistics*, 19:440–481, 2015b. ISSN 1262-3318. doi: 10.1051/ps/2014032. URL http://dx.doi.org/10.1051/PS/2014032.

Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

Rehan Ahmad and Kevin S Xu. Effects of contact network models on stochastic epidemic simulations. In *International Conference on Social Informatics*. Springer, 2017.

Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis*. Springer, Berlin, third edition, 2006. ISBN 978-3-540-32696-0; 3-540-32696-0.

Patrick Ango Nze. Critères d'ergodicité géométrique ou arithmétique de modèles linéaires perturbés à représentation markovienne. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(3):371–376, 1998. ISSN 0764-4442. doi: 10.1016/S0764-4442(97)82997-7. URL https://doi.org/10.1016/S0764-4442(97)82997-7.

Ernesto Araya and Yohann De Castro. Latent Distance Estimation for Random Geometric Graphs. In *Advances in Neural Information Processing Systems*, pages 8721–8731, 2019.

Miguel A. Arcones and Evarist Giné. Limit theorems for U-processes. *Ann. Probab.*, 21(3):1494–1542, 1993. ISSN 0091-1798. URL http://links.jstor.org/sici?sici=0091-1798(199307)21:3<1494:LTF>2.0.CO;2-I&origin=MSN.

Sylvain Arlot. Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3):1–106, 2019.

E Armengol et al. Evaluating link prediction on large graphs. In *Artificial intelligence research and development: proceedings of the 18th international conference of the Catalan association for artificial intelligence*, volume 277, 2015.

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.

Jushan Bai. Testing parametric conditional distributions of dynamic models. *The Review of Economics and Statistics*, 85(3):531–549, 2003.

Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 07 2016. doi: 10.1214/15-AOP1025. URL https://doi.org/10.1214/15-AOP1025.

Albert-László Barabási. Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413, 2009.

Jean-Marc Bardet, Paul Doukhan, Gabriel Lang, and Nicolas Ragache. Dependent Lindeberg central limit theorem and some applications. *ESAIM: Probability and Statistics*, 12:154–172, 2008.

Ludwig Baringhaus and Norbert Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988.

Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, Feb 2011. ISSN 0370-1573. doi: 10.1016/j.physrep.2010.11.002. URL http://dx.doi.org/10.1016/j.physrep.2010.11.002.

Ehrhard Behrends. *Introduction to Markov chains*. Advanced Lectures in Mathematics. Friedr. Vieweg & Sohn, Braunschweig, 2000. ISBN 3-528-06986-4. doi: 10.1007/978-3-322-90157-6. URL https://doi.org/10.1007/978-3-322-90157-6.

Jan Beirlant, László Györfi, and Gábor Lugosi. On the asymptotic normality of the L1- and L2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309 – 318, 12 2008.

P. Bertail and S. Clémençon. A renewal approach to Markovian U-statistics. *Math. Methods Statist.*, 20 (2):79–105, 2011. ISSN 1066-5307. doi: 10.3103/S1066530711020013. URL https://doi.org/10.3103/S1066530711020013.

Quentin Berthet and Nicolai Baldin. Statistical and Computational Rates in Graph Logistic Regression. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2719–2730. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/berthet20a.html.

R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer New York, 1996. ISBN 9780387948461. URL https://books.google.fr/books?id=F4hRy1F1M6QC.

Rajendra Bhatia and Ludwig Elsner. The Hoffman-Wielandt inequality in infinite dimensions. *Proceedings of the Indian Academy of Sciences, Mathematical Sciences, 104, pp. 483-494*, 104, 08 1994. doi: 10.1007/BF02867116.

Frank Bickenbach, Eckhardt Bode, et al. Markov or not Markov-this should be a question. Technical report, Kiel working paper, 2001.

A. Biswas and B. Biswas. Community-based link prediction. *Multimedia Tools and Applications*, 76, 2016.

Béla Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2001. doi: 10.1017/CBO9780511814068.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Nonbacktracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs. *The Annals of Probability*, 46(1):1 – 71, 2018. doi: 10.1214/16-AOP1142. URL https://doi.org/10.1214/16-AOP1142.

I. S. Borisov and N. V. Volodko. A note on exponential inequalities for the distribution tails of canonical von Mises' statistics of dependent observations. *Statist. Probab. Lett.*, 96:287–291, 2015. ISSN 0167-7152. doi: 10.1016/j.spl.2014.10.008. URL https://doi.org/10.1016/j.spl.2014.10.008.

Thomas R Boucher. A Hoeffding inequality for Markov chains using a generalized inverse. *Statistics & probability letters*, 79(8):1105–1107, 2009.

Solesne Bourguin, Charles-Philippe Diez, and Ciprian A Tudor. Limiting behavior of large correlated wishart matrices with chaotic entries. *Bernoulli*, 27(2):1077–1102, 2021.

Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *Probability Theory and Related Fields*, 178:1215–1289, 12 2020. doi: 10.1007/s00440-020-00998-3.

Heinz Breu and David G. Kirkpatrick. Unit disk graph recognition is NP-hard. *Computational Geometry*, 9(1):3 – 24, 1998. ISSN 0925-7721. doi: https://doi.org/10.1016/S0925-7721(97)00014-X. URL http://www.sciencedirect.com/science/article/pii/S092577219700014X. Special Issue on Geometric Representations of Graphs.

Zhan Bu, Yuyao Wang, Hui-Jia Li, Jiuchuan Jiang, Zhiang Wu, and Jie Cao. Link prediction in temporal networks: Integrating survival analysis and game theory. *Information Sciences*, 498, 2019.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012. doi: 10.1561/2200000024. URL https://doi.org/10.1561/2200000024.

Sébastien Bubeck and Shirshendu Ganguly. Entropic CLT and phase transition in high-dimensional Wishart matrices. *CoRR*, abs/1509.03258, 2015. URL http://arxiv.org/abs/1509.03258.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z. Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, Jan 2016. ISSN 1042-9832. doi: 10.1002/rsa.20633. URL http://dx.doi.org/10.1002/rsa.20633.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2(none):1153 – 1194, 2008. doi: 10.1214/08-EJS287. URL https://doi.org/10.1214/08-EJS287.

Cristina Butucea et al. Goodness-of-fit testing and quadratic functional estimation from indirect observations. *The Annals of Statistics*, 35(5):1907–1930, 2007.

Emmanuel Candes and Benjamin Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1):577–589, 2013.

Alain Celisse, Jean-Jacques Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6, 2012.

Saptarshi Chakraborty and Kshitij Khare. Consistent estimation of the spectrum of trace class data augmentation algorithms. *Bernoulli*, 25(4B):3832–3863, 2019.

Antoine Channarond. Random graph models: an overview of modeling approaches. *Journal de la Société Française de Statistique*, 156(3):56–94, 2015.

Moses Charikar, Sudipto Guha, Eva Tardos, and David Shmoys. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences*, 65, 2002.

Xiaming Chen and Yunwen Lei. Refined bounds for online pairwise learning algorithms. *Neurocomputing*, 275:2656–2665, 2018.

Yudong Chen and Jiaming Xu. Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices. *J. Mach. Learn. Res.*, 17:27:1–27:57, 2016.

Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423. PMLR, 2015.

Andreas Christmann and Ingo Steinwart. Support Vector Machines. *Support Vector Machines: Information Science and Statistics.*, 01 2008. doi: 10.1007/978-0-387-77242-4.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2606–2615. JMLR.org, 2016.

Gabriela Ciolek and Patrice Bertail. New Bernstein and Hoeffding type inequalities for regenerative Markov chains. *Latin American journal of probability and mathematical statistics*, 16:1–19, 02 2019. doi: 10.30757/ALEA.v16-09.

Stephan Clémençon, Patrice Bertail, and Gabriela Cioł ek. Statistical learning based on Markovian data maximal deviation inequalities and learning rates. *Ann. Math. Artif. Intell.*, 88(7):735–757, 2020. ISSN 1012-2443. doi: 10.1007/s10472-019-09670-6. URL https://doi.org/10.1007/s10472-019-09670-6.

Andrea E.F. Clementi, Francesco Pasquale, Angelo Monti, and Riccardo Silvestri. Information spreading in stationary Markovian evolving graphs. *2009 IEEE International Symposium on Parallel & Distributed Processing*, May 2009. doi: 10.1109/ipdps.2009.5160986. URL http://dx.doi.org/10.1109/IPDPS.2009.5160986.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-statistics. *Ann. Statist.*, 36(2):844–874, 04 2008. doi: 10.1214/009052607000000910. URL https://doi.org/10.1214/009052607000000910.

John B Conway. *A course in functional analysis*, volume 96. Springer, 2019.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013.

Jesper Dall and Michael Christensen. Random Geometric Graphs. *Phys. Rev. E*, 66:016121, Jul 2002. doi: 10.1103/PhysRevE.66.016121. URL https://link.aps.org/doi/10.1103/PhysRevE.66.016121.

Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. Community detection in sparse time-evolving graphs with a dynamical Bethe-Hessian. *Advances in Neural Information Processing Systems*, 33, 2020.

Lorenzo Dall'Amico and Romain Couillet. Community detection in sparse realistic graphs: Improving the bethe hessian. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2942–2946. IEEE, 2019.

Yohann De Castro, Claire Lacour, and Thanh Mai Pham Ngoc. Adaptive estimation of nonparametric geometric graphs. *Math. Stat. Learn.*, 2(3):217–274, 2019. ISSN 2520-2316.

Victor H. de la Peña and S. J. Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate U-statistics. *Ann. Probab.*, 23(2):806–816, 1995. ISSN 0091-1798. URL http://links.jstor.org/sici?sici=0091-1798(199504)23:2<806:DIFTTP>2.0.CO;2-H&origin=MSN.

Victor de la Pena and Evarist Giné. Decoupling, from dependence to independence, randomly stopped processes, U-statistics and processes, martingales and beyond. *Journal of the American Statistical Association*, 95, 09 2000. doi: 10.2307/2669501.

P. Del Moral and A. Guionnet. Central limit theorem for nonlinear filtering and interacting particle systems. *Ann. Appl. Probab.*, 9(2):275–297, 1999. ISSN 1050-5164. doi: 10.1214/aoap/1029962742. URL https://doi.org/10.1214/aoap/1029962742.

R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1993. ISBN 9783540506270. URL https://books.google.fr/books?id=cDqNW6k7_ZwC.

Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, 23(2):151–178, 2008.

Paul Doukhan and Marcel Ghindès. Estimations dans le processus "$X_{n+1} = f(X_n) + \varepsilon''_n$. *C. R. Acad. Sci. Paris Sér. A-B*, 291(1):61–64, 1980. ISSN 0151-0509.

Quentin Duchemin. Reliable Time Prediction in the Markov Stochastic Block Model. preprint, March 2022. URL https://hal.archives-ouvertes.fr/hal-02536727.

Quentin Duchemin and Yohann De Castro. Markov random geometric graph, MRGG: A growth model for temporal dynamic networks. *Electron. J. Stat.*, 16(1):671–699, 2022. doi: 10.1214/21-ejs1969. URL https://doi.org/10.1214/21-ejs1969.

Quentin Duchemin and Yohann De Castro. A new procedure for Selective Inference with the Generalized Linear Lasso. 2022. URL https://hal.archives-ouvertes.fr/hal-03622196.

Quentin Duchemin and Yohann De Castro. The Random Geometric Graph: Recent developments and perspectives. 2022. URL https://hal.archives-ouvertes.fr/hal-03622277.

Quentin Duchemin, Kangning Liu, Carlos Fernandez-Granda, and Jakob Assländer. Optimized dimensionality reduction for parameter estimation in MR fingerprinting via deep learning. *ISMRM*, 3750(1): 189–195, 2020.

Quentin Duchemin, Yohann De Castro, and Claire Lacour. Three rates of convergence or separation via U-statistics in a dependent framework. *JMLR*, 2022a. URL https://hal.archives-ouvertes.fr/hal-03603516.

Quentin Duchemin, Yohann De Castro, and Claire Lacour. Concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. *Bernoulli*, 2022b. URL https://hal.archives-ouvertes.fr/hal-03014763.

Daniel M Dunlavy, Tamara G Kolda, and Evrim Acar. Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5, 2011.

Daniele Durante and David Dunson. Bayesian logistic Gaussian process models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 194–201. PMLR, 2014.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Probability and moment inequalities for additive functionals of geometrically ergodic Markov chains. *arXiv preprint arXiv:2109.00331*, 2021.

Josep Díaz, Dieter Mitsche, and Xavier Pérez-Giménez. On the connectivity of dynamic random geometric graphs. pages 601–610, 01 2008.

P. Eichelsbacher and U. Schmock. Large deviations for products of empirical measures of dependent sequences. *Markov Process. Related Fields*, 7(3):435–468, 2001. ISSN 1024-2953.

Peter Eichelsbacher and Uwe Schmock. Rank-dependent moderate deviations of U-empirical measures in strong topologies. *Probab. Theory Related Fields*, 126(1):61–90, 2003. ISSN 0178-8051. doi: 10.1007/s00440-003-0254-6. URL https://doi.org/10.1007/s00440-003-0254-6.

Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's inequality for general Markov chains and its applications to statistical learning. *J. Mach. Learn. Res.*, 22(139):1–35, 2021. ISSN 1532-4435.

Yanqin Fan. Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis*, 62(1):36 – 63, 1997. ISSN 0047-259X. doi: https://doi.org/10.1006/jmva.1997.1672. URL http://www.sciencedirect.com/science/article/pii/S0047259X97916729.

Yanqin Fan and Aman Ullah. On goodness-of-fit tests for weakly dependent processes using kernel method. *Journal of Nonparametric Statistics*, 11(1-3):337–360, 1999.

Yingjie Fei and Yudong Chen. Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Transactions on Information Theory*, 65(1):551–571, 2018.

Xu Feng, JC Zhao, and Ke Xu. Link prediction in complex networks: a clustering perspective. *The European Physical Journal B*, 85, 2012.

Tamara Fernández and Arthur Gretton. A maximum-mean-discrepancy goodness-of-fit test for censored data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2966–2975, 2019.

Déborah Ferré, Loïc Hervé, and James Ledoux. Limit theorems for stationary Markov processes with L^2-spectral gap. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(2):396–423, 2012. ISSN 0246-0203. doi: 10.1214/11-AIHP413. URL https://doi.org/10.1214/11-AIHP413.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

G. Fort, E. Moulines, P. Priouret, and P. Vandekerkhove. A simple variance inequality for U-statistics of a Markov chain with applications. *Statist. Probab. Lett.*, 82(6):1193–1201, 2012. ISSN 0167-7152. doi: 10.1016/j.spl.2012.02.001. URL https://doi.org/10.1016/j.spl.2012.02.001.

Magalie Fromont and Béatrice Laurent. Adaptive goodness-of-fit tests in a density model. *Ann. Statist.*, 34(2):680–720, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000119. URL https://doi.org/10.1214/009053606000000119.

Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, Ulrike Von Luxburg, et al. Two-sample hypothesis testing for inhomogeneous random graphs. *Annals of Statistics*, 48(4):2208–2229, 2020.

E. N. Gilbert. Random Plane Networks. *Journal of the Society for Industrial and Applied Mathematics*, 9(4): 533–543, 1961. doi: 10.1137/0109045. URL https://doi.org/10.1137/0109045.

Pisier Gilles. *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1989.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016. ISBN 978-1-107-04316-9. doi: 10.1017/CBO9781107337862. URL https://doi.org/10.1017/CBO9781107337862.

Evarist Giné, RafałLatał a, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1, 2019.

Peter W. Glynn and Dirk Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statist. Probab. Lett.*, 56(2):143–146, 2002. ISSN 0167-7152. doi: 10.1016/S0167-7152(01)00158-4. URL https://doi.org/10.1016/S0167-7152(01)00158-4.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1292–1301. JMLR.org, 2017.

Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106, 2009.

Zheng-Chu Guo, Yiming Ying, and Ding-Xuan Zhou. Online regularized learning with pairwise loss functions. *Advances in Computational Mathematics*, 43(1):127–150, 2017.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165, 2014.

M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE Journal on Selected Areas in Communications*, 27(7):1029–1046, 2009. doi: 10.1109/JSAC.2009.090902.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Semidefinite programs for exact recovery of a hidden community, 2016.

Fang Han. An exponential inequality for U-statistics under mixing conditions. *J. Theoret. Probab.*, 31 (1):556–578, 2018. ISSN 0894-9840. doi: 10.1007/s10959-016-0722-4. URL https://doi.org/10.1007/s10959-016-0722-4.

Fang Han and Tianchen Qian. On inference validity of weighted U-statistics under data heterogeneity. *Electron. J. Stat.*, 12(2):2637–2708, 2018. doi: 10.1214/18-EJS1462. URL https://doi.org/10.1214/18-EJS1462.

Desmond Higham, Marija Rasajski, and Natasa Przulj. Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics (Oxford, England)*, 24:1093–9, 05 2008a. doi: 10.1093/bioinformatics/btn079.

Desmond J. Higham, Marija Rašajski, and Nataša Pržulj. Fitting a geometric graph to a protein–protein interaction network. *Bioinformatics*, 24(8):1093–1099, 03 2008b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn079.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. doi: 10.1198/016214502388618906. URL https://doi.org/10.1198/016214502388618906.

Steven C. H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9): 1–30, 2015.

Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.*, 2(2):85–114, 1993. ISSN 1066-5307.

Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

Emily M Jin, Michelle Girvan, and Mark EJ Newman. Structure of growing social networks. *Physical review E*, 64(4):046132, 2001.

Rong Jin, Shijun Wang, and Yang Zhou. Regularized Distance Metric Learning: Theory and Algorithm. In *NIPS*, volume 22, pages 862–870. Citeseer, 2009.

Emilien Joly and Gábor Lugosi. Robust estimation of U-statistics. *Stochastic Process. Appl.*, 126(12):3760–3773, 2016. ISSN 0304-4149. doi: 10.1016/j.spa.2016.04.021. URL https://doi.org/10.1016/j.spa.2016.04.021.

Galin L. Jones. On the Markov chain central limit theorem. *Probab. Surveys*, 1:299–320, 2004. doi: 10.1214/154957804100000051. URL https://doi.org/10.1214/154957804100000051.

Jonathan Jordan and Andrew R. Wade. Phase Transitions for Random Geometric Preferential Attachment Graphs. *Advances in Applied Probability*, 47(2):565–588, Jun 2015. ISSN 1475-6064. doi: 10.1239/aap/1435236988. URL http://dx.doi.org/10.1239/aap/1435236988.

Purushottam Kar, Bharath K. Sriperumbudur, Prateek Jain, and Harish C. Karnick. On the Generalization Ability of Online Learning Algorithms for Pairwise Loss Functions. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–441–III–449. JMLR.org, 2013.

Brian Karrer and M.E.J. Newman. Stochastic blockmodels and community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 83, 2011.

Nicolas Keriven and Samuel Vaiter. Sparse and smooth: Improved guarantees for spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 16(1):1330 – 1366, 2022. doi: 10.1214/22-EJS1986. URL https://doi.org/10.1214/22-EJS1986.

Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, Feb 2017. ISSN 0090-5364. doi: 10.1214/16-aos1454. URL http://dx.doi.org/10.1214/16-AOS1454.

Vladimir Koltchinskii and Evarist Giné. Random Matrix Approximation of Spectra of Integral Operators. *Bernoulli*, 6, 02 2000. doi: 10.2307/3318636.

Andrea Kölzsch, Erik Kleyheeg, Helmut Kruckenberg, Michael Kaatz, and Bernd Blasius. A periodic Markov model to formalize animal migration on a network. *Royal Society open science*, 5(6):180438, 2018.

H Kruckenberg, GJDM Müskens, and BS Ebbinge. Data from: A periodic Markov model to formalise animal migration on a network [white-fronted goose data], 2018. URL http://dx.doi.org/10.5441/001/1.kk38017f.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1312486110. URL https://www.pnas.org/content/110/52/20935.

Stanisł aw Kwapień and Wojbor A. Woyczyński. *Random series and stochastic integrals: single and multiple*. Probability and its Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. ISBN 0-8176-3572-6. doi: 10.1007/978-1-4612-0425-1. URL https://doi.org/10.1007/978-1-4612-0425-1.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, 44(3):907 – 927, 2016. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43, 2015.

Matthieu Lerasle, Nelo Molter Magalhães, and Patricia Reynaud-Bouret. Optimal kernel selection for density estimation. In *High dimensional probability VII*, volume 71 of *Progr. Probab.*, pages 425–460. Springer, [Cham], 2016. doi: 10.1007/978-3-319-40519-3_19. URL https://doi.org/10.1007/978-3-319-40519-3_19.

Levin. *Markov chains and mixing times*. American Mathematical Soc., 2017.

Fuchun Li and Greg Tkacz. A Consistent Bootstrap Test for Conditional Density Functions with Time-Dependent Data. Staff working papers, Bank of Canada, 2001.

Hua Liang and Pang Du. Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics*, 6:1838–1846, 2012.

Fredrik Lindsten, Randal Douc, and Eric Moulines. Uniform ergodicity of the particle Gibbs sampler. *Scand. J. Stat.*, 42(3):775–797, 2015. ISSN 0303-6898. doi: 10.1111/sjos.12136. URL https://doi.org/10.1111/sjos.12136.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284, 2016.

Siqi Liu, Sidhanth Mohanty, Tselil Schramm, and Elizabeth Yang. Testing thresholds for high-dimensional sparse random geometric graphs. *arXiv preprint arXiv:2111.11316*, 2021.

Caroline Lo, Justin Cheng, and Jure Leskovec. Understanding Online Collection Growth Over Time: A Case Study of Pinterest. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 545–554, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349147. doi: 10.1145/3041021.3054189. URL https://doi.org/10.1145/3041021.3054189.

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3.

Mathurin Massias, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Dual extrapolation for sparse generalized linear models. *Journal of Machine Learning Research*, 21(234):1–33, 2020.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4): 1119–1141, 2017.

Amit Meir and Mathias Drton. Tractable Post-Selection Maximum Likelihood Inference for the Lasso. *arXiv: Methodology*, 2017.

Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 02 1996. doi: 10.1214/aos/1033066201. URL https://doi.org/10.1214/aos/1033066201.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993. ISBN 3-540-19832-6. doi: 10.1007/978-1-4471-3267-7. URL https://doi.org/10.1007/978-1-4471-3267-7.

Claus Müller. *Analysis of spherical symmetries in Euclidean spaces*, volume 129. Springer Science & Business Media, 2012.

Pierre-Loïc Méliot. Asymptotic representation theory and the spectrum of a Random Geometric Graph on a compact Lie group. *Electron. J. Probab.*, 24:85 pp., 2019. doi: 10.1214/19-EJP305. URL https://doi.org/10.1214/19-EJP305.

Fragkiskos Papadopoulos, Maksim Kitsak, M.Ángeles Serrano, Marian Boguna, and Dmitri Krioukov. Popularity versus similarity in growing networks. *Nature*, 489(7417):537–540, Sep 2012. ISSN 1476-4687.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:no. 79, 32, 2015. doi: 10.1214/EJP.v20-4039. URL https://doi.org/10.1214/EJP.v20-4039.

Sandrine Péché and Vianney Perchet. Robustness of Community Detection to Random Geometric Perturbations. *Advances in Neural Information Processing Systems*, 33, 2020.

Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18, 2007.

Mathew Penrose. *Random Geometric Graphs*, volume 5. OUP Oxford, 2003.

Marianna Pensky and Teng Zhang. Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13, 2017.

Amelia Perry and Alexander S Wein. A semidefinite program for unbalanced multisection in the stochastic block model. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 64–67. IEEE, 2017.

Robert T. Powers and Erling Størmer. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1 – 33, 1970. doi: cmp/1103842028. URL https://doi.org/.

V. Preciado and A. Jadbabaie. Spectral analysis of virus spreading in random geometric networks. *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 4802–4807, 2009.

B. M. Pötscher. Effects of Model Selection on Inference. *Econometric Theory*, 7(2):163–185, 1991. ISSN 02664666, 14694360. URL http://www.jstor.org/stable/3532042.

Yimo Qin, Bin Zou, Jingjing Zeng, Zhifei Sheng, and Lei Yin. Online regularized pairwise learning with non-iid observations. *International Journal of Wavelets, Multiresolution and Information Processing*, page 2150041, 2021.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71, 2004. doi: 10.1214/154957804100000024. URL https://doi.org/10.1214/154957804100000024.

R.T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970. ISBN 9780691015866. URL https://books.google.fr/books?id=1TiOka9bx3sC.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(2), 2010.

Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: a survey. *ACM Computing Surveys (CSUR)*, 51(2):1–37, 2018.

Rimantas Rudzkis and Aleksej Bakshaev. Goodness of fit tests based on kernel density estimators. *Informatica*, 24(3):447–460, 2013.

Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000. ISSN 0091-1798. doi: 10.1214/aop/1019160125. URL https://doi.org/10.1214/aop/1019160125.

Glen A. Satten, Maiying Kong, and Somnath Datta. Multisample adjusted U-statistics that account for confounding covariates. *Stat. Med.*, 37(23):3357–3372, 2018. ISSN 0277-6715. doi: 10.1002/sim.7825. URL https://doi.org/10.1002/sim.7825.

Jeffrey Scargle. Publication bias: the "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, 14:91–106, 01 2000.

René Schott and G Stacey Staples. Dynamic Geometric Graph Processes : Adjacency Operator Approach. *Advances in applied Clifford algebras*, 20(3):893–921, 2010.

Oleksandr Shchur and Stephan Günnemann. Overlapping community detection with graph neural networks. *Deep Learning on Graphs Workshop, KDD*, 2019.

Yandi Shen, Fang Han, and Daniela Witten. Exponential inequalities for dependent V-statistics via random Fourier features. *Electron. J. Probab.*, 25:1–18, 2020. doi: 10.1214/20-ejp411. URL https://doi.org/10.1214/20-ejp411.

Xiang-yu Shi, Bo Liang, and Qi Zhang. Post-selection inference of generalized linear models based on the Lasso and the elastic net. *Communications in Statistics - Theory and Methods*, 0(0):1–18, 2020. doi: 10.1080/03610926.2020.1821892. URL https://doi.org/10.1080/03610926.2020.1821892.

Grace S. Shieh, Richard A. Johnson, and Edward W. Frees. Testing independence of bivariate circular data and weighted degenerate U-statistics. *Statist. Sinica*, 4(2):729–747, 1994. ISSN 1017-0405.

Steve Smale and Ding-Xuan Zhou. Online learning with Markov sampling. *Analysis and Applications*, 7 (01):87–113, 2009.

Anna L. Smith, Dena M. Asta, and Catherine A. Calder. The geometry of continuous latent space models for network data. *Statist. Sci.*, 34(3):428–453, 08 2019. doi: 10.1214/19-STS702. URL https://doi.org/10.1214/19-STS702.

Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. *arXiv preprint arXiv:1905.10826*, 2019.

Marc A Suchard, Robert E Weiss, and Janet S Sinsheimer. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular biology and evolution*, 18(6):1001–1013, 2001.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019. doi: 10.1073/pnas.1810420116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1810420116.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2018. ISBN 978-0-262-03924-6.

Minh Tang, Daniel L. Sussman, and Carey E. Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, Jun 2013. ISSN 0090-5364. doi: 10.1214/13-aos1112. URL http://dx.doi.org/10.1214/13-AOS1112.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018. doi: https://doi.org/10.1002/cjs.11313. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11313.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46, 07 2015. doi: 10.1214/17-AOS1564.

Xiaoying Tian and Jonathan Taylor. Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44(2):480–499, 2017. doi: https://doi.org/10.1111/sjos.12261. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12261.

Xiaoying Tian, Joshua R Loftus, and Jonathan E Taylor. Selective inference with unknown variance via the square-root Lasso. *Biometrika*, 105(4):755–768, 09 2018.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(none):1456 – 1490, 2013. doi: 10.1214/13-EJS815. URL https://doi.org/10.1214/13-EJS815.

Ryan J. Tibshirani, Alessandro Rinaldo, Rob Tibshirani, and Larry Wasserman. Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46(3):1255–1287, 2018. ISSN 00905364, 21688966. URL https://www.jstor.org/stable/26542824.

Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012.

Samuel Vaiter, Mohammad Golbabaee, Jalal Fadili, and Gabriel Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287, 2015.

Sara A Van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

A. Van Der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, 2013. ISBN 9781475725452.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Gang Wang and Zun Lin. On the performance of multi-message algebraic gossip algorithms in dynamic Random Geometric Graphs. *IEEE Communications Letters*, PP:1–1, 07 2014. doi: 10.1109/LCOMM.2014.2344047.

Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization Bounds for Online Learning Algorithms with Pairwise Loss Functions. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 13.1–13.22, Edinburgh, Scotland, 2012. PMLR. URL https://proceedings.mlr.press/v23/wang12.html.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684):440–442, 1998.

Changshuai Wei, Robert C. Elston, and Qing Lu. A weighted U statistic for association analyses considering genetic heterogeneity. *Stat. Med.*, 35(16):2802–2814, 2016. ISSN 0277-6715. doi: 10.1002/sim.6877. URL https://doi.org/10.1002/sim.6877.

Roi Weiss and Boaz Nadler. Learning parametric-output hmms with two aliased states. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 635–644, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/weiss15.html.

Zheng Xie and Tim Rogers. Scale-invariant geometric random graphs. *Physical Review E*, 93(3), Mar 2016. ISSN 2470-0053. doi: 10.1103/physreve.93.032310. URL http://dx.doi.org/10.1103/PhysRevE.93.032310.

Zheng Xie, Jiang Zhu, Dexing Kong, and Jianping Li. A Random Geometric Graph built on a time-varying Riemannian manifold. *Physica A: Statistical Mechanics and its Applications*, 436:492 – 498, 2015. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2015.05.076. URL http://www.sciencedirect.com/science/article/pii/S0378437115004914.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15:521–528, 2002.

Jie Xu, Yuan Yan Tang, Bin Zou, Zongben Xu, Luoqing Li, and Yang Lu. The generalization ability of online SVM classification based on Markov sampling. *IEEE transactions on neural networks and learning systems*, 26(3):628–639, 2014.

Kevin S. Xu. Stochastic block transition models for dynamic networks. *CoRR*, 2014.

Kevin S. Xu and Alfred O. Hero. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562, 2014. doi: 10.1109/JSTSP.2014.2310294.

Sikun Yang and Heinz Koeppl. Dependent relational gamma process models for longitudinal networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5551–5560. PMLR, 10–15 Jul 2018.

Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, and Rong Jin. Detecting communities and their evolutions in dynamic social networks - a bayesian approach. *Machine Learning*, 82, 2011.

Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 587–594, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581644. doi: 10.1145/1390334.1390435. URL https://doi.org/10.1145/1390334.1390435.

Yiming Ying and Ding-Xuan Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 04 2014. ISSN 0006-3444. doi: 10.1093/biomet/asv008. URL https://doi.org/10.1093/biomet/asv008.

Jingjing Zeng, Bin Zou, Yimo Qin, Qian Chen, Jie Xu, Lei Yin, and Hongwei Jiang. Generalization ability of online pairwise support vector machine. *Journal of Mathematical Analysis and Applications*, 497(2):124914, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *International Conference on Computational Learning Theory*, pages 173–187. Springer, 2005.

Xiaoxia Zhang, Quentin Duchemin, Kangning Liu, Sebastian Flassbeck, Cem Gultekin, Carlos Fernandez-Granda, and Jakob Assländer. Cramér-Rao bound-informed training of neural networks for quantitative MRI. *Magnetic Resonance in Medicine*, 2022. doi: https://doi.org/10.1002/mrm.29206. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.29206.

Peilin Zhao, Steven C. H. Hoi, Rong Jin, and Tianbao Yang. Online AUC maximization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 233–240, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Bin Zou, Hai Zhang, and Zongben Xu. Learning from uniformly ergodic Markov chains. *Journal of Complexity*, 25(2):188–200, 2009.

Konstantin Zuev, Marian Boguna, Ginestra Bianconi, and Dmitri Krioukov. Emergence of soft communities from Geometric Preferential Attachment. *Scientific reports*, 5, 01 2015. doi: 10.1038/srep09421.

Dynamique de croissance de grands réseaux à l'aide de chaînes de Markov cachées

Résumé :

La première partie de cette thèse vise à introduire de nouveaux modèles de graphes aléatoires rendant compte de l'évolution temporelle des réseaux. Plus précisément, nous nous concentrons sur des modèles de croissance où à chaque instant un nouveau noeud s'ajoute au graphe existant selon une dynamique markovienne latente. Nous nous intéresserons particulièrement au Stochastic Block Model et aux Graphes Aléatoires Géométriques pour lesquels nous proposons des algorithmes permettant d'estimer les paramètres du modèle et de résoure des problèmes de prédiction de lien ou de filtrage collaboratif.

L'étude théorique des algorithmes précédemment décrits mobilisent des résultats probabilistes poussés. Nous avons notamment dû recourir à une inégalité de concentration pour les U-statistiques d'ordre deux pour des chaînes de Markov uniformément ergodiques. La deuxième partie de cette thèse présente la preuve de ce résultat ainsi que certaines applications importantes en Statistiques.

Toujours motivés par des problèmes de prédictions liens dans les graphes, nous nous intéressons dans un dernier chapitre aux procédures d'inférence post-sélection dans le cadre de la régression logistique avec pénalité L^1. Nous prouvons un théorème central limite sous la distribution conditionnelle à l'événement de sélection et nous en déduisons des procédures de test et des intervalles de confiance asymptotiquement valides.

Mots clefs : Graphes aléatoires, Chaines de Markov, Estimation non-paramétrique, Concentration de la mesure, Opérateurs intégraux, Inférence post-sélection, Apprentissage séquentiel.

Growth dynamics of large networks using hidden Markov chains

Abstract :

The first part of this thesis aims at introducing new models of random graphs that account for the temporal evolution of networks. More precisely, we focus on growth models where at each instant a new node is added to the existing graph according to some latent Markovian dynamic. We are particularly interested in the Stochastic Block Model and in Random Geometric Graphs for which we propose algorithms to estimate the unknown parameters or functions defining the model and to solve link prediction problems.

The theoretical analysis of the above-mentioned algorithms requires advanced probabilistic tools. In particular, we needed a concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. The second part this thesis is dedicated to the proof of this result and to its important consequences in Statistics.

Still motivated by link prediction problems in graphs, we study post-selection inference procedures in the framework of logistic regression with L^1 penalty. We prove a central limit theorem under the distribution conditional on the selection event and derive asymptotically valid testing procedures and confidence intervals.

Keywords: Random Graphs, Markov chains, Non-Parametric Estimation, Measure Concentration, Integral Operators, Post-Selection Inference, Online Learning.

PARIS EST SUP ÉCOLE DOCTORALE
Mathématiques et STIC

Université Gustave Eiffel