Kernel spectral learning and inference in random geometric graphs

Ernesto Araya Valdivia

Thèse de doctorat

NNT: 2020UPASM020

Kernel spectral learning and inference in random geometric graphs

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées
Unité de recherche: Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.
Référent: : Faculté des sciences d'Orsay

Thèse présentée et soutenue en visioconférence totale, le 16/12/2020, par

Ernesto ARAYA VALDIVIA

Composition du jury:

Christophe GIRAUD Professeur, Université Paris-Saclay	Président
Bruno PELLETIER Professeur, Université Rennes 2	Rapporteur et examinateur
Viet Chi TRAN Professeur, Université Gustave Eiffel	Rapporteur et examinateur
Olga KLOPP Professeur, ESSEC Business School	Examinatrice
Pierre LATOUCHE Professeur, Université de Paris	Examinateur
Yohann DE CASTRO Professeur, École Centrale de Lyon	Directeur de thèse

Remerciements

Mes premiers remerciements s'adressent à Yohann, mon directeur de thèse, pour ces trois années de travail ensemble, pour m'avoir proposé de travailler sur le sujet de graphes géométriques qui me passionne, et pour sa sympathie qui a fait que notre collaboration soit très agréable. Merci encore pour les conseils et pour m'avoir montré des techniques mathématiques, ainsi que les *tricks of the trade* des statisticiens.

Je voudrais aussi remercier à mon jury, pour avoir accepté être là, dans ces conditions très spéciales. Merci pour ces commentaires, leur avis, et pour la discussion scientifique qui j'ai vraiment adorée. Une mention particulière à mes deux rapporteurs Bruno et Chi pour avoir lu mon manuscrit en avant-première.

Un grand merci à ma famille pour avoir été toujours là et pour leur support inconditionnel à la distance. À Eric et Willy pour avoir un grand coeur, et pour m'avoir accueilli de la meilleure façon dans sa maison. À mes amis chiliens, rencontrés à Paris, pour tant de moments agréables et pour leur soutien jour à jour. À mes amis au Chili, pour déjà presque une vie d'amitié. À tous les collègues à Orsay pour les discussions passionnants de maths, de littérature, enfin de la vie.

À la fin, j'ai aussi une pensée à tous mes mort, qui sont toujours dans mes souvenirs.

Contents

List of Figures

Chapter 1
Introduction

Contents

Cette thèse présente une ensemble de contributions au problème d'inférence statistique en graphes aléatoires géométriques, ainsi qu'au problème de concentration du spectre des matrices à noyau. Cette classe de matrices a des applications dans l'ensemble des méthodes à noyaux en apprentissage statistique, ainsi que dans la théorie des grands réseaux dans le régime dense. Elles seront un des objets centraux d'étude dans l'ensemble de ce mémoire.

Le chapitre 2 s'intéresse au problème de concentration des matrices à noyau, plus spécifiquement la concentration relative de ses valeurs propres. Les chapitres 3 et 4 portent sur l'inférence de l'information latente dans le modèle de réseaux géométrique, qui est l'un de plus utilisés parmi les modèles de graphes aléatoires à espace latent. Dans le cas du chapitre 3 on se concentre en graphes représentés par la sphère Euclidienne et au chapitre 4 on se place dans le cas de la boule Euclidienne. Même si les deux modèles ont des similarités formelles, ils ont aussi des différences intéressantes concernant la distribution des degrés des graphes qu'ils génèrent. Chaque chapitre correspond à un article soumis où en cours de soumission pour publication. L'article associé au chapitre 2 est dans le deuxième tours de révision et l'article du chapitre 3 est publié dans les *proceedings* de la conference *NeurIPS* 2019, qui a eu lieu en décembre 2019 à Vancouver, Canada.

En grande partie, la recherche présentée ici est motivée par l'omniprésence des réseaux au niveau de modélisation des systèmes complexes et pour les questions mathématiques qui émergent lorsqu'on essaie d'extraire de l'information sur eux, ou de comprendre les algorithmes qui fonctionnent sur des réseaux massifs, comme les réseaux de communication ou les réseaux sociaux. Nous allons étudier des questions théoriques liées à ces problématiques.

On se concentrera sur un type particulier de réseaux dense: ceux qui sont représentables par des noyaux qu'on appellera *graphons.* La représentation d'un graphe, objet éminemment discret, par une fonction va nous permettre d'utiliser des outils d'analyse fonctionnelle et harmonique, qui vont nous révéler l'information essentielle sur le graphe, ou sur la classe à laquelle il appartient.

Dans le chapitre 2, intitulé “Relative concentration of random kernel matrices”, nous allons nous concentrer sur le modèle des matrices aléatoires à noyau, qui sont centrales dans les méthodes à noyaux en apprentissage statistique et qui servent aussi comme des outils pour l’étude des réseaux denses. Nous montrerons des bornes de concentration pour le spectre de ces matrices, sous conditions de régularité liés à la vitesse de convergence de l’expansion spectrale des noyaux respectifs. Notre méthode permet d’obtenir des bornes plus fines pour la fluctuation des valeurs propres individuels que celles dans la plupart de la littérature et dans un cadre plus général. En particulier, on évite l’hypothèse de positivité du noyaux, ce qui est fondamental pour l’étude des réseaux, où les noyaux associés sont souvent non-positifs. Nous montrerons que ce cadre-ci est bien adapté au cas des noyaux invariants par rotation définis sur la sphère Euclidienne, qui sont reliés au modèle des graphes géométriques. Le matériel ici présenté est basé sur [Araya 2020].

Le chapitre 3, qui a pour titre “Latent distances estimation for RRG on the sphère", est consacré à l’étude du modèle des graphes aléatoires géométriques sur la sphère Euclidienne unitaire. Dans ce modèle, chaque noeud d’un graphe est aléatoirement place sur la sphère et un arc est formé entre deux noeuds avec une probabilité qui dépend de la distance entre les noeuds (codifiée par son produit scalaire) et un noyau de connexion. On s’intéresse au problème d’estimation des distances latentes entre les noeuds, à partir de la seule observation de la matrice d’adjacence d’un graphe généré par le modèle des graphes géométriques. On propose un algorithme pour la estimation de la matrice des distances, qui reçois comme des entrées la matrice d’adjacence et la dimension de la sphère latente. On montre des garanties théoriques pour l’erreur de estimation, de type Frobenius, en grande probabilité. On assume de conditions sur le trou spectral d’un opérateur intégral associé au noyau de connexion. La méthode proposée est fondée sur les propriétés de concentration du spectre du noyaux invariants par rotation sur la sphère. Cet algorithme est utilisé comme sous-routine pour la estimation de la dimension, lorsque elle est inconnue. Nous renforçons l’analyse théorique par des expériences numériques et simulations. Le matériel ici présenté est basé sur [Araya 2019].

Dans le chapitre 4, intitulé “Random geometric graphs on the Euclidean ball”, nous allons étudier le modèle de graphes aléatoires géométriques avec la boule Euclidienne unitaire comme espace sous-jacent. De façon similaire au chapitre 3 nous allons considérer de graphes représentables par des noyaux qui dépendent seulement du produit scalaire. Le fait que les points dans la boule ont un degré de liberté de plus par rapport à la sphère (où la norme est fixé et égal à 1) donne au modèle sur la boule plus de flexibilité au niveau de modélisation. En particulier, nous montrerons que pour certaines noyaux de connexion la distribution de la séquence de degrés, dans le graphe généré, est similaire à une loi de puissance, qui est suivant mentionnée comme une des distributions prépondérantes dans la modélisation des réseaux réels. D’autre part nous allons étudier deux problèmes de estimation sur ce modèle: l’estimation de la norme des points latents et l’estimation des distances latentes. Ceci étendre les idées développées dans le chapitre 3. Nous illustrons les méthodes développées par des expériences numériques.

1.0.1 Notation

On utilisera tout au long de cette thèse les notations suivantes. Soient f et g deux fonctions réelles. On dit que $f(x) \lesssim g(x)$, si et seulement s'il existe $C > 0$ réel, tel que $f(x) \leq Cg(x)$. De façon similaire, on dit que $f(x) \lesssim_\alpha g(x)$, pour $\alpha \in \mathbb{R}$ pour renforcer que C peut dépendre de α. On utilisera la notation asymptotique comme de manière usuelle, c'est-à-dire, on dit que $f(x) = \mathcal{O}(g(x))$ si et seulement s'il existe $N \in \mathbb{R}$ tel que $f(x) \lesssim g(x)$, pour $|x| > N$. De manière analogue, on dit que $f(x) = \mathcal{O}_\alpha(g(x))$ s'il existe $N \in \mathbb{R}$ tel que $f(x) \lesssim_\alpha g(x)$, pour $|x| > N$. Tant pour une matrice que pour un opérateur compact dans un espace d'Hilbert on notera $\|\cdot\|_{op}$ la norme d'opérateur, qui correspond à la plus grande valeur singulière.

1.1 Graphes aléatoires et le modèle du graphon

Dans les chapitres 3 et 4 on étudiera problèmes d'inférence sur des réseaux, qui l'on supposera générés par des modèles de graphes aléatoires à espace latent. Tous les modèles qui seront étudiés sont représentables par des noyaux grâce au formalisme du graphon.

Les graphons (contraction de *graph* et *functions*) sont des noyaux symétriques bornés qui jouent un rôle fondamental dans la définition du modèle W-*random graph* et aussi dans la théorie de limites (lorsque sa taille tend vers l'infini) de graphes denses. Le modèle W-random graph a été introduit par Diaconis and Freedman [Diaconis 1981] dans les années 80, mais sa popularité a augmenté au cours de la dernière décennie suite au développement de la théorie des limites des graphes, avec les travaux fondateurs de Lóvasz et collaborateurs [Lovász 2006b, Lovász 2006a, Borgs 2008, Borgs 2012, Borgs 2010]. Comme nous allons le voir, cette théorie donne un cadre assez général pour la modélisation de graphes aléatoires et permet d'utiliser des outils puissants provenant de l'analyse fonctionnelle et harmonique pour en déduire des propriétés combinatoires.

Étant donné un espace mesurable (Ω, μ), un graphon en Ω est une fonction $W : [0,1] \times [0,1] \to [0,1]$ symétrique et mesurable. Il est possible d'utiliser [Lovasz 2012][Chap.11], sans perte de généralité, l'espace $\Omega = [0,1]$ et μ la mesure de Lebesgue, mais dans cette thèse nous allons préférer la définition plus générale.

Pour obtenir un graphe simple à partir d'un graphon W en (Ω, μ), nous opérons de la façon suivante. D'abord on considère l'échantillon des points, qui seront les noeuds dans le graphe généré, $\{X_i\}_{i\in[n]}$ en Ω selon la loi μ. Ensuite, on construit la matrice suivante

$$\Theta_{ij} := W(X_i, X_j)$$

qu'on appelle la *matrice de probabilités*. Étant donné Θ, on définit la matrice d'adjacence A comme une matrice aléatoire symétrique avec des entrées A_{ij} i.i.d pour $i < j$ avec une loi Bernoulli qui satisfait

$$\mathbb{P}(A_{ij} = 1 | X_1 \cdots, X_n) = W(X_i, X_j)$$

les entrées de la diagonale sont toutes zéros (restriction de graphe simple). Plusieurs modèles classiques de graphes aléatoires peuvent être exprimés, dans le cas dense, par des graphons. Un des premiers modèles de graphes aléatoires a été introduit par Erdös et Rényi (et porte ses noms) où deux noeuds quelconques sont connectés de façon indépendante avec la même probabilité de connexion $p \in [0, 1]$. Celui-ci correspond au modèle W-random graph avec graphon constant égal à p. De la même façon, on peut voir que le modèle stochastique par blocs (SBM) appartient à la classe des modèles W-random graph en considérant une partition mesurable de Ω, qu'on appelle $\{\Omega_i\}_{i\in[K]}$ (ici on a K communautés), et un graphon défini par $W_{SBM}(x, y) = p_{ij}$ pour $(x, y) \in \Omega_i \times \Omega_j$, où $p_ij \in [0, 1]$ est la probabilité de connection entre membres de la communauté i avec membres de la communauté j. Le modèle de graphe aléatoire géométrique classique, introduit en [Gilbert 1961], où les noeuds sont placés dans un espace métrique et il existe un arc entre deux noeuds s'ils sont assez proches, est aussi représentable par un graphon. Prenons par exemple où l'espace ambiant est le cube $\Omega = [0, 1]^d$ avec la mesure uniforme et le graphon défini par $W_g(x, y) = \mathbb{1}_{\|x-y\|\leq\tau}$. Dans ce cas, le modèle W-random graph est équivalent au modèle de graphe aléatoire géométrique qui connecte les points plus proches qu'un seuil $\tau > 0$.

À part son intérêt dans le cadre de la modélisation, les graphons représentent des limites de séquences des graphes denses. Le sens précis de cette convergence est donné par la *cut distance* [Lovasz 2012], qui est une distance dans l'espace de graphons qui le rend un espace compact [Lovasz 2012]. Tout graphe fini a une représentation par un graphon et la convergence d'une séquence des graphes sera équivalente à la convergence des graphons dans le sens de la cut distance. La convergence dans le sens de la cut distance est aussi reliée à la convergence des homéomorphismes ou sous-motifs donnés par des graphes finis fixés. Nous n'utiliserons pas, de façon directe, la cut distance dans ce manuscrit. Il important de mentionner la notion d'isomorphisme faible: deux graphons W_1 et W_2 sont faiblement isomorphes s'il existe deux transformations ψ_1 et ψ_2 en S_Ω telles que $W_1(\psi_1(x), \psi_1(y)) = W_2(\psi_2(x), \psi_2(y))$, où S_Ω est l'ensemble des transformations qui préservent la mesure μ. Être faiblement isomorphe est une relation d'équivalence dans l'espace de graphons. Dans le contexte de problèmes statiques sur des graphons, cette propriété est liée à des problèmes d'identifiabilité, car un graphon W quelconque définit le même modèle W-random graph que W^ψ, pour tout $\psi \in S_\Omega$, où $W^\psi(x, y) = W(\psi(x), \psi(y))$. Fréquemment en problèmes d'inférence on considère une des classes d'équivalences définies para la notion d'isomorphisme faible.

En dépit de l'importance de la cut distance dans la théorie des graphes, la plupart des travaux dans le domaine des statistiques utilisent d'autres normes (l'article [Klopp 2017b] est une des exceptions) sur l'espace de graphons, plus classiques dans le contexte d'analyse fonctionnelle comme les normes L^p. Le fait que les normes L^p majorent la *cut norm* (utilisé pour définir la cut distance), permet d'en déduire des propriétés intéressantes. Dans cette thèse on s'occupera principalement des normes du type L^2. Dans la section suivante nous allons définir le spectre du graphon qui sera central pour nos algorithmes d'inférence d'information latente.

1.2 Concentration du spectre des matrices à noyau

Les matrices à noyau jouent un rôle fondamental dans une grande variété des méthodes en apprentissage statistique, comme l'analyse de composantes principales, la régression de crête et, surtout dans la dernière décennie, l'analyse des réseaux denses. Souvent son spectre est utilisé dans des algorithmes d'apprentissage et, par conséquent, des bornes pour les valeurs propres sont nécessaires pour avoir des garanties théoriques sur l'erreur de ces méthodes.

Une matrice de noyau de taille $n \times n$ à des entrées de la forme $K(X_i, X_j)$ où $K : \Omega \times \Omega \to \mathbb{R}$ est un noyau et $\{X_i\}_{i\in[n]} \subset \Omega$ un ensemble de points. On supposera que $\{X_i\}_{i\in[n]}$ son tirés de façon i.i.d avec une loi commun μ et que K est une fonction $L^2(\mu \times \mu)$ symétrique. On voit en particulier que la matrice de probabilités Θ de la section précédente est une matrice à noyau.

On considérera plutôt la normalisation suivante $(T_n)_{ij} := \frac{1}{n}K(X_i, X_j)$. À chaque noyau on associe un opérateur intégral défini par $T_K : L^2(\mu) \to L^2(\mu)$

$$T_K f(x) = \int_\Omega K(x, y)f(y)d\mu(y)$$

Étant donné que K est de carré intégrable, on sait que T_K est un opérateur compact, par un résultat classique d'analyse fonctionnelle, et alors son spectre $\lambda(T_K)$ est un ensemble énumérable qui a 0 comme seul point d'accumulation. On peut identifier l'ensemble $\lambda(T_K)$ avec une séquence en $\mathbb{R}^\mathbb{N}$ qui à un limite égale à 0. On considère l'indexation $|\lambda_1(T_K)| \geq |\lambda_2(T_K)| \cdots$ où les $\lambda_i(T_K)$ sont les valeurs propres de T_K. Le fait que $K \in L^2$ implique que l'opérateur T_K est dans la classe des opérateurs d'Hilbert-Schmidt, c'est-à-dire, on a $\sum_{i\geq 1} \lambda_i^2 < \infty$.

Pour deux séquences de carrés sommables $a, b \in \mathbb{R}^\mathbb{N}$, on définit la distance δ_2 par

$$\delta_2(a, b) = \inf_{\pi\in\Pi} \sqrt{\sum_{i\geq 1} \left(a_i - b_\pi(i)\right)^2}$$

où Π est l'ensemble des permutations sur $\mathbb{N}$ avec support fini. Pour tout $n \in \mathbb{N}$, on identifiera une suite $a \in \mathbb{R}^{[n]}$ avec son extension $\tilde{a} \in \mathbb{R}^\mathbb{N}$ obtenue en rajoutant un nombre infini des zéros, ce qui permet de comparer le spectre des objets de dimension finie (matrices) avec le spectre des opérateurs.

Koltchinskii et Giné [Koltchinskii 2000] ont montré que le spectre de $(\tilde{T}_n)_{ij} = (1 - \delta_{ij})(T_n)_{ij}$ converge vers le spectre de T_K, dans le sens de la distance $\delta_2(\cdot, \cdot)$.

Théorème 1. *[Koltchinskii 2000, Thm.3.1] Si $\int K^2(x, y)d\mu(x)d\mu(y) < \infty$ alors*

$$\delta_2(\lambda(\tilde{T}_n), \lambda(T_W)) \to 0$$

lorsque $n \to \infty$.

Celui-ci représente la loi de grands nombres pour le spectre des opérateurs de carré intégrable. Dans [Koltchinskii 2000], les auteurs prouvent aussi un théorème central limite pour cette convergence.

Le théorème spectral pour des opérateurs compacts donne l'expansion suivante pour le noyau

$$K(x,y) \overset{L^2}{=} \sum_{i\geq 1} \lambda_i \phi_i(x)\phi_i(y)$$

où $\phi_i(\cdot)$ sont les fonctions propres de l'opérateur intégral (c'est-à-dire $T_K\phi_i = \lambda_i\phi_i$), qui forment une base Hilbertienne de $L^2(\mu)$. Dans le cadre d'estimation de la fonction graphon en [De Castro 2020], des inégalités de concentration ont été montrés pour $\delta_2(T_W, \frac{1}{n}\Theta)$ où W est un graphon. Dans ce travail, les auteurs se placent dans le cas des espaces symétriques compacts et, parmi eux, la sphère euclidienne $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ que l'on étudiera de plus près pour le modèle de graphes géométriques dans la section 2.7 et le chapitre 3. Comme exemples des résultats obtenus en [De Castro 2020] dans le cas sphérique pour des graphons qui satisfassent des conditions de régularité du type Sobolev[1] (relié à la vitesse à laquelle les valeurs propres convergeant vers 0), on compte le théorème suivant.

Théorème 2. *[De Castro 2020] Soit W un graphon sur $\mathbb{S}^{d-1}$ de la forme $W(x,y) = f(\langle x,y\rangle)$, où f appartient à un espace de Sobolev avec poids $Z^s_{w_\gamma}\big((-1,1)\big)$ alors on a pour n suffisamment grand*

$$\delta_2(\lambda\big(T_n\big), \lambda(T_W)) \lesssim_\alpha C\Big(\frac{\log n}{n}\Big)^{\frac{s}{2s+d-1}}$$

avec probabilité plus grande que $1-\alpha$.

Une borne sur la distance $\delta_2(\cdot,\cdot)$ implique de façon immédiate des inégalités pour des valeurs propres individuelles, tout simplement car $|\lambda_i(T_n) - \lambda_i(T_W)| \leq \delta_2\big(\lambda_i(T_n), \lambda_i(T_W)\big)$. Pour des noyaux positives, inégalités de la forme $|\lambda_i(T_n) - \lambda_i(T_W)| = \mathcal{O}(\frac{1}{\sqrt{n}})$ ont été montres, en utilisant des techniques des espaces d'Hilbert à noyau reproductif [Rosasco 2010]. Un chemin différent est adopté en [Belkin 2018], où des techniques de théorie de l'approximation sont utilisés pour montrer une inégalité pour le spectre des matrices à noyaux radiaux positifs et suffisamment différentiables en $\mathbb{R}^d$. Plus précisément, ils considèrent des noyaux de la forme $K(x,y) = f(\|x-y\|)$, où f est une fonction réelle, qui satisfait $|\frac{d^l}{dt^l}f(t)| \leq l!M^l$ pour l suffisamment grand. Ils obtiennent

Théorème 3. *[Belkin 2018, Thm.2] Pour un noyau K positif suffisamment differentiable on a pour tout $1 \leq i \leq n$*

$$|\lambda_i(T_W) - \lambda_i(T_n)| \lesssim \exp\left(-ci^{1/d}\right)$$

Un avantage du théorème 3 par rapport à $|\lambda_i(T_n) - \lambda_i(T_W)| = \mathcal{O}(\frac{1}{\sqrt{n}})$, c'est la dépendance en i à droite dans l'inégalité. Cela permettre d'obtenir des résultats plus précis pour les valeurs propres plus petits, comme a été montré en [Braun 2006].

[1] La définition formelle sera donnée dans le chapitre 2.

Une façon générale d'obtenir des bornes, pour $|\lambda_i(T_W) - \lambda_i(T_n)|$, avec une dépendance en i est d'utiliser des théorèmes de perturbation relative pour matrices [Ipsen 1998]. Ceci est un sujet bien connu dans le domaine de l'analyse numérique dont une grande précision est nécessaire. Si A et B sont deux matrices, qu'on supposera réelles et symétriques, l'idée est de changer la quantité $|\lambda_i(A) - \lambda_i(B)|$ pour une des quantités suivantes

$$\frac{|\lambda_i(A) - \lambda_i(B)|}{|\lambda_i(A)|}, \quad \frac{|\lambda_i(A) - \lambda_i(B)|}{\sqrt{|\lambda_i(A)||\lambda_i(B)|}} \quad \text{ou} \quad \frac{|\lambda_i(A) - \lambda_i(B)|}{(|\lambda_i(A)|^p + |\lambda_i(B)|^p)^{\frac{1}{p}}}$$

pour $p \geq 1$, lorsque $\lambda_i(A)$ et $\lambda_i(B)$ sont différents de zéro. Bien que les deux derniers ont des bonnes propriétés (comme la symétrie par rapport à A et B), on préférera la première, c'est-à-dire $\frac{|\lambda_i(A)-\lambda_i(B)|}{|\lambda_i(A)|}$. La raison est que dans notre contexte B sera une matrice aléatoire, la matrice T_n pour être plus précis, et A sera une approximation (déterministe) de dimension finie de l'opérateur T_W. Cela permettre d'avoir seulement des termes déterministes à droite de l'inégalité. Le but est d'obtenir inégalités du type

$$|\lambda_i(T_n) - \lambda_i(T_W)| \lesssim |\lambda_i(T_W)| n^{-h} \tag{1.1}$$

avec grande probabilité, avec $h \in (0, 1)$. On appelle les résultats de la forme (1.1), inégalités de type Weyl.

Des inégalités du type (1.1) ont été montrés en [Braun 2006], où ils sont aussi appelés *scaling bounds*, car le terme $|\lambda_i(T_W)|$ permet de "mettre à l'échelle" le côté droit. Plus spécifiquement, ils montrent le théorème suivant

Théorème 4. *[Braun 2006, Thm .4 et Sec. 6.1] Supposons que K est un noyau semi-définie positive, continu et borné avec valeurs propres $\lambda_1, \lambda_2, \cdots$. Alors, pour toute $1 \leq R \leq n$ et $\alpha \in (0,1)$, on a avec probabilité plus grande que $1 - \alpha$*

$$|\lambda_i(T_n) - \lambda_i| = \mathcal{O}(\lambda_i R \sqrt{\frac{\log R}{\lambda_R^2 n}} + \Lambda_{>R} + \sqrt{\frac{\Lambda_{>R}}{n}}) \tag{1.2}$$

ou $\Lambda_{>R} = \sum_{k>R} \lambda_k$. En outre, si $\lambda_i = \mathcal{O}(i^{-\delta})$ avec $\delta > 1$ on a pour tout $1 \leq i \leq n$

$$|\lambda_i(T_n) - \lambda_i| \leq \mathcal{O}(n^{-\frac{\delta-1}{2+3\delta}}) \tag{1.3}$$

Si $\lambda_i = \mathcal{O}(e^{-\delta i})$, alors $|\lambda_i(T_n) - \lambda_i| = \mathcal{O}_\alpha(n^{-\frac{1}{3}} \log^2 n)$.

Bien qu'étant une borne relative, l'équation (1.2) a des inconvénients. Pour un i fixé, le terme à droite ne tends pas vers 0 lorsque $n \to \infty$ (il le fait si $R \to \infty$), ce qui est donné par la présence du terme de biais $\Lambda_{>R}$. Cela est remédié pour le cas $\lambda_i = \mathcal{O}(i^{-\delta})$ en (1.3). Aussi, dans (1.2) la présence du terme $\lambda_R^{-1/2}$ la rend peu pratique pour le cas quand λ_R es petit. D'autre part, l'inégalité (1.3) implique que la vitesse de convergence est plus lente que $\mathcal{O}(\frac{1}{\sqrt{n}})$, qu'on peut obtenir (au moins dans le cas positive) par une méthode basée sur des espaces d'Hilbert à noyau reproductif (RKHS) [Rosasco 2010, Blanchard 2007].

Dans l'esprit d'obtenir une vitesse de convergence comparable au théorème 3 et dans un cadre plus général, avec des conditions de décroissance pour les valeurs propres analogues au théorème 4, on a montré le théorème suivant.

Théorème 5. *[Araya 2020] Soit W un noyau avec valeurs propres $\lambda_1, \lambda_2, \cdots$ et fonctions propres correspondant $\phi_1, \phi_2, \cdots$,. On suppose que i es tel que $\lambda_i \neq 0$ et que $\| \sum_{i=1}^n |\lambda_i| \phi_i^2 \|_\infty \leq \infty$, alors il existe $n_0(i) \in \mathbb{N}$ tel que pour tout $n \geq n_0$ on a*

$$|\lambda_i(T_n) - \lambda_i| = \mathcal{O}_\alpha\Big(\lambda_i \sqrt{\frac{\mathcal{V}_1(R(i))}{n}}\Big) \tag{1.4}$$

où $\mathcal{V}_1(i) = \| \sum_{k=1}^i \phi_k^2 \|_\infty$ et $R(i) := \min\big\{R \in \mathbb{N} : |\lambda_i| > \sum_{k>R} |\lambda_k| \vee \sqrt{R \sum_{k>R} \lambda_k^2}\big\}$. En outre, on suppose que $\lambda_i = \mathcal{O}(i^{-\delta})$ et $\|\phi_i\|_\infty = \mathcal{O}(i^s)$. Alors, avec probabilité plus grand que $1 - \alpha$ on a, pour $s \geq 1$

$$|\lambda_i(T_n) - \lambda_i| = \begin{cases} \mathcal{O}_\alpha\Big(i^{-\delta + \frac{\delta}{\delta-1}(s+\frac{1}{2})} n^{-\frac{1}{2}}\Big) & si \quad 1 \leq i \leq n^{\frac{\delta-1}{\delta}\frac{1}{2s+1}} \\ \mathcal{O}_\alpha\Big(i^{-\delta + 1 + \frac{\delta}{\delta-1}(s+\frac{1}{2})} n^{-\frac{1}{2}}\Big) & si \quad n^{\frac{\delta-1}{\delta}\frac{1}{2s+1}} \leq i \leq n^{\frac{1}{2s}} \\ \mathcal{O}_\alpha\Big(i^{-\delta+s+1} n^{-\frac{1}{2}}\Big) & si \quad n^{\frac{1}{2s}} \leq i \leq n \end{cases}$$

Si on suppose que $\lambda_i = \mathcal{O}(e^{-\delta i})$ et $\|\phi_i\|_\infty = \mathcal{O}(e^{is})$, alors pour $s \geq 1$

$$|\lambda_i(T_n) - \lambda_i| = \begin{cases} \mathcal{O}_\alpha\Big(e^{-\delta i + (s+\frac{1}{2})\log i} n^{-\frac{1}{2}}\Big) & si \quad 1 \leq i \leq n^{\frac{1}{2s}} \\ \mathcal{O}_\alpha\Big(e^{-\delta i + s \log i} n^{-\frac{1}{2}}\Big) & si \quad n^{\frac{1}{2s}} \leq i \leq n \end{cases}$$

Pour la preuve du théorème 5, on s'est inspiré des travaux [Koltchinskii 2000, Braun 2006, De Castro 2020] et on a suivi la stratégie suivante: on introduit une approximation W_R de rang R du noyau W, et on considère la matrice à noyau $T_{n,R} = W_R(X_i, X_j)$ avec le même échantillon X_i qui définit T_n. Ensuite, on peut voir T_n comme une perturbation de $T_{n,R}$ et en utilisant des résultats de perturbation relative on arrive à montrer que $|\lambda_i(T_n) - \lambda_i(T_{n,R})| \leq |\lambda_i| \|A\|_{op}$ et $|\lambda_i(T_{n,R}) - \lambda_i| \leq |\lambda_i| \|B\|_{op}$, où A et B sont deux matrices aléatoires. La matrice B es de la forme $\Phi_R \Phi_R^T - \mathrm{Id}_R$ où Φ_R es une matrice de $R \times n$ avec colonne i égale à $(\phi_1(X_i), \cdots, \phi_R(X_i))$. On obtient des bornes pour $\|B\|_{op}$ en utilisant des théorèmes de concentration pour matrices, comme le théorème de Bernstein matriciel [Tropp 2012, Thm.6.1] (où alternativement [Vershynin 2012a, Thm.5.1]). Le terme $\|A\|_{op}$ est plus difficile à borner et on utilise une majoration par la norme de Frobenius, laquelle on borne par à l'aide des inégalités de concentration pour U-statistiques [Gine 2000].

Pour résumer, nos principales contributions au corpus d'inégalités de concentration du type relatif pour des matrices à noyau sont l'obtention des inégalités du type Weyl avec des vitesses de convergences égales ou meilleures que celles déjà présentes dans la littérature, et dans un cadre moins restreint(sans la hypothèse contraignante que T_W est positif). De plus, on applique les résultats au cadre de noyaux qui dépendent seulement du produit scalaire et on montre les liens avec les graphes géométriques sur la sphère.

1.3 Inférence des graphes aléatoires géométriques

L'inférence des graphes aléatoires est une famille de problèmes liés à la récupération d'information d'un modèle de graphes (paramétrique ou non-paramétrique) à partir de données (ici, on supposera qu'on dispose d'une et seulement une observation du graphe). Une sous-classe de problèmes est liée à la détection de la présence de sous-structure dans le graphe et à la décision de quel modèle parmi deux (voire plus) est plus vraisemblable compte tenues des données. Ces problèmes sont formalisés, du point de vue statistique, comme des problèmes de test d'hypothèse [Arias-Castro 2014, Arias-Castro 2015]. Un autre groupe de problèmes d'inférence concerne l'estimation d'un modèle ou des sous-structures cachées. Un des exemples les plus connus est le problème de détection communauté, où il existe une partition cachée des noeuds qui détermine la connexion entre eux (typiquement la connexion inter-communauté est plus forte que la connexion intra-communauté), comme dans le modèle SBM. L'objectif est de récupérer la partition cachée [Abbe 2017].

Ici, nous abordons le problème d'estimation d'information latente dans le cadre de graphes aléatoires géométriques. Nous allons considérer des graphes générés à partir d'un modèle W-random graph, avec graphon $W(x, y) = f(\langle x, y\rangle)$, où $f : [-1, 1] \to [0, 1]$ est appelée *fonction de connection*. Dans le chapitre 3, nous prendrons $\Omega = \mathbb{S}^{d-1}$ et on supposera que l'échantillon $\{X_i\}_{1\leq i\leq n}$ suit une loi uniforme sur la sphère, tandis que dans le chapitre 4 on considère $\Omega = \mathbb{B}^d := \{x \in \mathbb{R}^d : \|x\| = 1\}$ et μ appartiendra à une famille de mesures avec symétrie sphérique. Le modèle ici décrit représente une généralisation du modèle géométrique classique, associé au graphon $W_g(x, y) = \mathbb{1}_{\langle x,y\rangle>\tau}$. D'autres noms ont été considérés pour cette généralisation dans la littérature, comme graphes géométriques non-paramétriques [De Castro 2020] ou *soft random geometric graph* [Penrose 2016].

La question de détection de géométrie est abordée dans le cadre du test d'hypothèse dans [Bubeck 2016], où les auteurs montrent que si d (la dimension de la sphère) est négligeable devant n^3, alors il est possible de distinguer le modèle W_g de l'hypothèse nulle qui est le modèle d'Erdös-Rényi. Plus précisément, ils montrent qu'il est possible d'utiliser une quantité reliée au nombre de triangles comme statistique pour décider le test. Dans le cas où la dimension est grande par rapport à n^3, ils montrent que il est impossible de distinguer entre les deux modèles, car la distance de variation totale entre les deux distributions converge vers 0. Ceci constitue un exemple du phénomène de transition de phase. Plus de précisions sur cette transition son données dans [Rácz 2019].

Ici on étudie un problème de récupération: à partir de l'observation de la matrice d'adjacence, l'objectif est de trouver les distances latentes (pour le cas sphérique il suffit de trouver $\langle X_i, X_j\rangle$). La récupération des distances latentes a des applications dans le domaine de localisation des senseurs [Li 2009, Eren 2017] et la prédiction de liens dans les réseaux sociaux [Sarkar 2010].

On suit une approche spectrale, qui a des similarités avec celui étudié dans [Sussman 2014], dans le cadre des graphes RDPG (random dot product graphs). D'autre part, les techniques qu'on utilise sont à la base de certaines méthodes d'apprentissage de variétés de faible dimension [Levin 2017].

Dans notre méthode l'idée est d'utiliser des éléments d'analyse harmonique sur la sphère pour trouver une décomposition convenable de la fonction f. Dans ce contexte, il est connu qu'il existe $\{\phi_{l,k}\}_{l\geq 0;1\leq k\leq d_l}$ une base Hilbertienne[2] de $L^2(\mathbb{S}^{d-1})$ telle que pour tout W de la forme $W(x,y) = f(\langle x,y\rangle)$ on a $T_W\phi_{l,k} = \lambda_l\phi_{l,k}$, où $\lambda_l \in \mathbb{R}$. Chaque $\{\phi_{l,k}\}$ est un polynôme sur la sphère de degré l que l'on les appelle les harmoniques de la sphère. L'opérateur projection de $L^2(\mathbb{S}^{d-1})$ dans l'espace de polynômes de degré l sur la sphère est un opérateur à noyau et son noyau a une formule explicite en fonction des polynômes de Gegenbauer du degré correspondant. Dans le cas de polynômes linéaires, cela implique la formule suivante

$$\langle x,y\rangle = \frac{1}{d}\sum_{j=1}^{d}\phi_{1,j}(x)\phi_{1,j}(y) \tag{1.5}$$

De façon équivalente, on a $\mathcal{G}^* = \Phi\Phi^T$ où Φ est une matrice de taille $n \times d$ qui a pour lignes $(\phi_{1,1}(X_i),\cdots,\phi_{1,d}(X_i))$ pour $1 \leq i \leq n$.

On propose un algorithme, qu'on appelle HEiC(Harmonic EigenCluster) qui reçois comme entrée une matrice d'adjacence et la dimension de la sphère d, et sa sortie est $\hat{\mathcal{G}}$, un estimateur de $\mathcal{G}^*$. L'algorithme trouve un cluster de taille exactement d valeurs propres proches entre eux et calcule $\hat{\mathcal{G}} := \frac{1}{d}\sum_{i=1}^{d}\hat{v}_i\hat{v}_i^T$, où $\hat{v}_1,\cdots,\hat{v}_d$ sont les vecteurs propres associes. On montre que sous une condition de trou spectral pour le spectre de T_W, l'algorithme a de garanties pour l'erreur en norme Frobenius avec grande probabilité. Le trou spectral est défini par

$$\Delta^* = \inf_{k\neq 1}|\lambda_k - \lambda_1|$$

Théorème 6. *Supposons que $W(x,y) = f(\langle x,y\rangle)$ et f appartient à un espace de Sobolev $Z^s([-1,1])$ et qui satisfait la condition de trou spectrale $\Delta^* > 0$, alors l'algorithme HEiC donne $\hat{\mathcal{G}}$ qui satisfait*

$$\|\mathcal{G}^* - \hat{\mathcal{G}}\|_F = \mathcal{O}(\Delta^{*-1}n^{-\frac{s}{2s+d-1}})$$

avec grand probabilité.

On montre aussi, que l'algorithme HEiC peut être utilisé comme sous-routine dans un algorithme qu'estime la dimension de la sphère, lorsque elle n'est pas reçue comme entrée. Plus précisément, on définit un algorithme HEiC-dim qui reçois comme entrée la matrice d'adjacence et $\mathcal{D}$ un ensemble de candidats pour la dimension et assigne un *score* à chaque candidat basé sur l'application de HEiC (le *score* correspond à la valeur empirique de Δ^*). La sortie est le candidat avec le plus grand *score*. On montre que si la condition de trou spectral est satisfaite, alors l'algorithme HEiC-dim trouve la bonne dimension pourvu que la vraie dimension est dans l'ensemble des candidats et que n est suffisamment grand. Comme la plupart des algorithmes spectraux, HEiC a une complexité en temps de l'ordre $\mathcal{O}(n^3)$.

[2] La séquence $\{d_l\}_{l\geq 0}$ es strictement croissante et satisfait $d_1 = d$.

D'autre part l'algorithme HEiC-dim a une complexité en temps $\mathcal{O}(n^3 \times |\mathcal{D}|)$, où $|\mathcal{D}|$ est la cardinalité de l'ensemble des candidats.

Dans le chapitre 4, on étudie le modèle de graphes géométriques dans la boule Euclidienne, c'est-à-dire on considère l'espace $\Omega = \mathbb{B}^d$ et un graphon de la forme $W(x, y) = f(\langle x, y\rangle)$. On considère une famille de mesures $\mathcal{F} = \{F_\nu\}_{\nu>1/2}$ avec symétrie sphérique, où chaque F_ν a la densité suivante (par rapport à la mesure de Lebesgue)

$$dF_\nu(x) = (1 - \|x\|^2)^{\nu-1/2}$$

Le choix de cette famille a deux raisons: la premiere est qu'elle est suffisamment riche pour obtenir une grande variété de profils de degrés (distribution de la séquence de degré) pour les graphes générés. La deuxième raison est technique, l'analyse harmonique sur la boule donné des résultats analogues au cas sphérique pour la famille $\mathcal{F}$.

Une des contraintes dans le modèle sphérique est le fait que tous les noeuds ont en moyenne le même degré (on peut montrer aussi que pour chaque noeud la distribution de son degré est très concentré par rapport à sa moyenne). Le fait que dans la boule il y a des points avec différente norme (et aussi qu'on a plus de contrôle sur la distribution avec la famille $\mathcal{F}$), permettent d'obtenir un profil des degrés avec un support plus étendu. En particulier, on montre que pour certaines fonctions de connexion et certaines mesures dans $\mathcal{F}$, il est possible d'obtenir distributions des degrés similaires à une lois de puissance décalée. De nombreuses études empiriques suggèrent qu'une bonne partie des réseaux qui apparaissent dans la pratique ont des lois de puissance pour ces degrés [Clauset 2009, Mitzenmacher 2003], ce qui rendre le modèle dans la boule attractif pour des applications.

On étudie des questions d'inférence dans ce modèle. En premier lieu, on prend le cas $W_g(x, y) = \mathbb{1}_{\langle x,y\rangle>\tau}$ et on montre que il n'est pas possible d'estimer τ et le paramètre ν de la mesure, en même temps, avec la seule information de la matrice d'ajacence. Ensuite, on montre que si on fixe la mesure et τ, il est possible d'estimer les normes latentes, c'est-à-dire, pour chaque i on définit ζ_i, un estimateur de $\|X_i\|$, et on montre que $\zeta_i \to \|X_i\|$ au sens presque sure. On étudie aussi le cas de τ inconnu.

Pour les distances latentes, on montre qu'il est possible d'étendre la méthode développée pour le cas sphérique, étant donné les similarités entre les décompositions spectrales dans le deux cas. En effet, dans la cas de la boule, il existe une base de $L^2(\mathbb{B}^d)$, constituée de polynômes qui sont des fonctions propres pour tout opérateur intégral avec noyau de la forme $W(x, y) = f(\langle x, y\rangle)$. De manière analogue au cas sphérique, le noyau de l'opérateur projection de L^2 dans l'espace des polynômes de degré 1 a une formule close similaire à eq.(1.5). On construit un estimateur basé sur les vecteurs propres de la matrice d'adjacence, comme dans le cas sphérique.

En résume, notre principale contribution au problème d'estimation des distances latentes pour un graphe géométrique en $\mathbb{S}^{d-1}$, est de proposer un algorithme efficient avec des garanties théoriques pour l'erreur, sous une supposition de type trou spectral. Utilisé comme sous-routine, l'algorithme permet de déterminer la dimension latente entre un ensemble de candidats.

Dans le cas de graphes géométriques en $\mathbb{B}^d$ nos principales contributions sont: montrer que le modèle est plus flexible en termes de distribution des degrés, en comparaison au modèle en $\mathbb{S}^{d-1}$. En particulier, on montre que pour certaines fonctions de connexion, le modèle présente des distributions des dégrés similaire à la loi de puissances. On construit des estimateurs pour les normes et les distances latentes et on montre qu'ils convergent vers la vraie valeur (sous certaines conditions de régularité).

Pour chaque algorithme et pour chaque estimateur, nous présentons une série des simulations et des expériences numériques qui permettent de vérifier les résultats empiriquement et qui suggèrent, à l'occasion, des extensions possibles de nos résultats.

Chapter 2

Relative concentration of random Kernel Matrices

Contents

2.1 Introduction

This chapter is devoted to the study of the concentration properties for the spectrum of a class of random matrices, known as *Kernel matrices*. Such matrices play an important role in the family of kernel methods, which are ubiquitous in machine learning, theory and applications alike [Hofmann 2008]. Important examples of the applications of kernel matrices are dimensionality reduction (Kernel PCA, for instance [Blanchard 2007]), sample covariance estimation [Koltchinskii 2017] and more recently in the analysis of dense networks [Klopp 2017a], data privacy [Kasiviswanathan 2015] and deep learning [Cao 2019].

It is well known [Koltchinskii 2000, Thm. 3.1] that each eigenvalue of such matrices converge to the eigenvalue of an associated kernel integral operator. The methods we present in this section, which are developed in detail in the paper [Araya 2020], offers a finite sample approach that quantify the eigenvalue convergence. Apart from the typical $\mathcal{O}(1/\sqrt{n})$ that comes from the use of concentration inequalities, the rates we present here have a scaling term, which allow us to obtain rates that are better than parametric and often exponential, which is a phenomenon that has already been observed, with more restrictive hypothesis and using a different approach, in [Belkin 2018].

The core of the work [Araya 2020] consists in two theorems, one dealing with a more general situation and with milder hypothesis and other under regularity hypothesis related with the decrease rate of the eigenvalues. Those hypothesis, which are common in the kernel in the literature and satisfied for a large class of kernel functions, will allow us to obtain more specialized results depending on the eigenvalue behavior of different kernel families. We apply our results to the case of *dot product kernels* defined on the Euclidean sphere $\mathbb{S}^{d-1}$, which is a family of rotation invariant kernels with the remarkable property that the associated integral operator can be diagonalized in the basis of spherical harmonics. The eigenfunction basis is fixed in this case (acting as a Fourier basis) and our results simplify, as the regularity will depend only on the eigenvalue decay rate. We highlight the connection of this type of kernels and the class of dense random geometric graphs, via the graphon formalism developed in [Lovász 2006b, Lovász 2006a, Borgs 2008, Borgs 2012, Borgs 2010]. This class of kernels has also been used recently in the context of neural networks and deep learning [Cao 2019].

2.1.1 Kernels, matrices and integral operators

Through this section, we will consider a probability space (Ω, μ) and a kernel will be a bivariate symmetric measurable function $K : \Omega \times \Omega \to \mathbb{R}$. We will assume, here and thereafter, that all the kernels are square integrable. Otherwise stated, they belong to $L^2(\Omega^2, \mu \times \mu)$, where $\mu \times \mu$ is product measure of μ with itself. Given the kernel K, we construct the integral operator $T_K : L^2(\Omega, \mu) \times L^2(\Omega, \mu) \to \mathbb{R}$, defined by

$$T_K f(x) = \int_\Omega f(y)K(x,y)d\mu(y)$$

where x is in Ω and $f \in L^2(\Omega, \mu)$. As usual, the space $L^2(\Omega, \mu)$ is endowed by the product $\langle f, g\rangle_{L^2} = \int_\Omega f(x)g(x)d\mu(x)$, which make it a Hilbert space. Given that we assume the square integrability of K, it is well known that T_K is in fact a compact self adjoint Hilbert-Schmidt operator [Hirsch 1999, p.216], hence by the spectral theorem for compact self adjoint operators (see Theorem 1 below) its spectrum is a real discrete (countable) set whose only accumulation point is 0. In that sense, a self adjoint compact operator can be regarded as an infinite dimensional analog of a (finite) symmetric matrix: both can be diagonalized and the spectrum is composed by real eigenvalues only. We recall that an eigenvalue of an operator is defined in an analogous manner to those of matrices, that is $\lambda \in \mathbb{R}$ is an eigenvalue of T_K if and only if there exists $\phi \in L^2(\Omega, \mu)$ such that the following relation holds

$$T_K\phi(x) = \lambda\phi(x)$$

for every $x \in \Omega$. In this context, the function ϕ is known as an eigenfunction associated to the eigenvalue λ, which we will assume to be normalized by $\|\phi\|_2 = 1$, where $\|\cdot\|_2$ is the norm induced by the inner product $\langle\cdot,\cdot\rangle_{L^2}$. In general one eigenvalue might be associated to more than one eigenfunction, defining the linear space E_λ, which is the subspace of $L^2(\Omega, \mu)$ that contains all the eigenfunctions associated with λ. The linear dimension of E_λ is known as the multiplicity of λ.

We recall the spectral theorem for compact self adjoint operators in the Hilbert space version, which will be used in the sequel

Theorem 1 (Spectral Theorem)**.** *Let T be a compact operator from a Hilbert space H to itself. For every eigenvalue λ, let E_λ denote the eigenspace associated to λ. Then the following holds*

- *The set $\lambda(T)$ of the eigenvalues of T is an infinite countable and bounded subset of $\mathbb{R}$.*
- *For all $\lambda \neq 0$ we have* $\dim(E_\lambda) \leq \infty$
- *For all λ, λ' in $\lambda(T)$ with $\lambda \neq \lambda'$ the spaces E_λ and $\mathbb{E}_{\lambda'}$ are orthogonal, meaning that $\langle f, g\rangle_{L^2} = 0$ for $f \in E_\lambda$ and $g \in E_{\lambda'}$.*
- *We have the following decomposition, in the $L^2(\Omega, \mu)$ sense*

$$T = \sum_{\lambda \in \lambda(T)\setminus\{0\}} \lambda P_\lambda$$

where P_λ is the orthogonal projector onto E_λ for $\lambda \neq 0$.

The previous theorem is one of the many incarnations of a classic statement. Its proof can be found, for example in [Hirsch 1999, Chap.6]. The following corollary will also be useful

Corollary 2. *The Hilbert space H has a countable Hilbert basis that consist of eigenfunctions $\{\phi_k\}_{k\in\mathbb{N}}$ of the operator T, where each ϕ_k with $k \in \mathbb{N}$ is associated with an eigenvalue $\lambda_k \neq 0$. In addition, the sequence of eigenvalues satisfy $\lambda_k \to 0$ when $k \to \infty$ and the following decomposition holds in the L^2 sense*

$$Tf = \sum_{k\in\mathbb{N}} \lambda_k \langle f, \phi_k\rangle_{L^2} \phi_k$$

As $K(x, \cdot)$ is itself a $L^2(\Omega)$ function, the following expansion holds in the L^2 sense

$$K = \sum_{k \in \mathbb{N}} \lambda_k \phi_k \otimes \phi_k \tag{2.1}$$

where $f \otimes g$ represents the tensor product between for $f, g \in L^2$, which is defined as $f \otimes g(x, y) = f(x)g(y)$.

From Corollary 2 we deduce that the set of eigenvalues $\lambda(T_K)$ can be identify with an element of c_0: the space of real sequences converging to 0. We will use the same notation $\lambda(T_K)$ for the abstract set of eigenvalues and for the sequence in c_0 (which definition is in use should be clear from the context). We will consider that in the sequence $\lambda(T_K)$ every eigenvalue appears as many times as its multiplicity, unless the contrary is stated. This correspond to the notion of *extended enumeration* defined in [Rosasco 2010, Sec. 2.5] (see [Kato 1995] for an earlier reference). Since every nonnegative sequence converging to zero can be rearranged in decreasing order, we will consider in the following the indexation for the eigenvalues $\{\lambda_k\}_{k \in \mathbb{N}}$:

$$|\lambda_0| \geq |\lambda_1| \geq \cdots 0$$

Given the eigenfunction normalization we have the following

$$\|K\|_{L^2} = \sum_{k \in \mathbb{N}} \lambda_k^2$$

where $\|K\|_{L^2}$ is the norm in $L^2(\Omega^2, \mu^2)$. As we have assumed that $\|K\|_{L^2} \leq \infty$, we see that the operator T_f is in fact a Hilbert-Schmidt operator. Also, its easy to see that given that we assume that $|\lambda_k|$ are ordered decreasingly, we have that $|\lambda_k| < C/\sqrt{k}$ for some constant C (which depends on K[1]). This will not be enough for our purposes and additional assumptions will be required.

The kernel matrix, which can be thought as a finite dimensional version of the integral operator T_K and is defined by sampling in the following manner: we start with the set of random variables $\{X_i\}_{\{1 \leq i \leq n\}} \in \Omega$, which are independent and identically distributed with law μ and we construct the matrix of pairwise evaluations

$$(K_n)_{ij} = K(X_i, X_j)$$

for $i, j \in [n]$. By construction, the kernel matrix is symmetric, so its spectrum $\lambda(K_n)$ lies in $\mathbb{R}$[2]. The cardinality of $\lambda(K_n)$ is n, but in order to compare the spectrum of K_n with the spectrum of T_K we will use the embedding of finite set of real numbers into the space c_0, which is simply defined by completing a sequence of finite lenght with zeros. We will use this identification in the sequel. Similar to the case of $\lambda(T_K)$, we will choose the indexation in the decreasing order in the absolute value for $\lambda(K_n)$.

Our goal is to prove relative concentration inequalities for eigenvalues of K_n with respect to the eigenvalues T_K. Asymptotic results are well known since the work of

[1] We can use here C=$\|K\|_{L^2}$, but this bound will not be used in the sequel.

[2] We use the notation $\lambda(\cdot)$ for matrices analogously to the operator case

Koltchniskii and Giné [Koltchinskii 2000], where the authors prove the convergence of $\lambda(\frac{1}{n}K_n)$ to $\lambda(T_K)$ in a ℓ_2-type norm, called δ_2 norm. They also provide a CLT [Koltchinskii 2000, Cor. 5.9] describing the asymptotic normality of the fluctuations of $\lambda(\frac{1}{n}K_n)$. Their proof includes the use of a truncation technique, common in functional analysis, which we will also use later. Notice that in the asymptotic results suggest that the correct normalization in this regime is $\frac{1}{n}$, which is a difference with the high-dimensional regime (see [El Karoui 2010]). In the sequel, we concentrate in the normalized version, defined by

$$T_n := \frac{1}{n}K_n$$

For any L^2 kernel K, where the expansion (2.1) holds, we define its truncation at level R, for any $R \in \mathbb{N}$ as follows

$$K_R = \sum_{k=1}^{R} \lambda_k \phi_k \otimes \phi_k$$

If for a given kernel K, there exists $R \in \mathbb{N}$ such that $K = K_R$, that is K coincides with its rank R decomposition, we will say that K is of finite rank (or, more specifically, of rank R). Otherwise, we say that K has infinite rank. We can see K_R as a finite rank approximation of K and the idea is to quantify this approximation under regularity conditions on K. This idea will be formalized in Section 2.4. Before we discuss the state of the art in kernel matrix concentration.

2.2 Kernel matrix concentration

The study of concentration inequalities in the context of matrix quantities deals mostly with the concentration of the their eigenvalues. Most of the inequalities in the literature deals with quantities of the form $|\|A\|_{op} - \mathbb{E}\|A\|_{op}|$ or $\|A - \mathbb{E}A\|_{op}$, where in the second case the matrix expectation must be understood entrywise and the operator norm corresponds to the largest singular value of a matrix or operator.

The results and techniques used to attack this problem vary drastically depending on the specific distributional assumptions (or lack thereof) for the entries of A. For instance, in the case of matrices with independent entries the operator norm concentration is well understood. Indeed in [Bandeira 2016] sharp inequalities are provided for this case. Following [El Karoui 2010], the kernel matrices concentration literature can be divided in two regimes: the low and the high dimensional cases. In the low dimensional case, the space Ω where the X_i's are defined is fixed (it is not intrinsically "low dimensional"). The high dimensional case deals with the case where Ω changes. Most often, $\Omega = \mathbb{R}^d$ and, in the high dimensional case, the most typical assumption is that $d/n \to \gamma$, where $\gamma \in (0, 1)$ is fixed. In the latter case, there are many remarkable asymptotic (the Marchenko-Pastur limit theorem for the spectral density [Marchenko 1967], for instance) and non-asymptotic results. We refer to the interested reader to [El Karoui 2010] for an exhaustive study of the concentration in the high dimensional case. From now on, we only deal with the low dimensional case only.

We can further divide the matrix concentration inequalities in two groups: the absolute concentration inequalities, on one hand, and the relative concentration inequalities, on the other. This categories derive from the types of measures used for quantifying matrix perturbations, which can either relative or absolute. This distinction is most common in the fields of linear algebra and numerical analysis (see [Ipsen 1998]), but they are becoming increasingly popular in statistics and machine learning. In the case of individual eigenvalues the most typical absolute measure is $|\lambda_i(A) - \lambda_i(B)|$ where A and B are two matrices (or operators). Some corresponding relative measures, which are common in the literature, are

$$\frac{|\lambda_i(A) - \lambda_i(B)|}{|\lambda_i(A)|}, \quad \frac{|\lambda_i(A) - \lambda_i(B)|}{\sqrt{|\lambda_i(A)||\lambda_i(B)|}} \quad \text{and} \quad \frac{|\lambda_i(A) - \lambda_i(B)|}{\left(|\lambda_i(A)|^p + |\lambda_i(B)|^p\right)^{\frac{1}{p}}}$$

for $p \geq 1$. We call a bound for any of this quantities of the *Weyl-type*, given the classic Weyl perturbation theorem [Bathia 1997][Cor.III.2.6]. Inequalities that involve the sum of eigenvalues are often called of the Hoffman-Weilandt type (also because of classic eigenvalue inequality with that name). On most occasions, the use relative eigenvalue inequalities results in better accuracy compared to the corresponding absolute inequalities.

Some of works that obtain absolute concentration inequalities in the context of kernel matrices are [Shawe-Taylor 2005],[Blanchard 2007],[Kasiviswanathan 2015] ,[De Castro 2020], [Amini 2020]. The relative approach has been considered in [Braun 2006] and, most recently, in [Belkin 2018] where an approximation theoretic method is directly employed (without using concentration). Even if the context is slightly different, the literature on concentration of sample covariance matrices (operators) is also relevant, since Gram type matrices are one of the prototypical examples of kernel matrices. Some examples are the works in [Vershynin 2012a] [Koltchinskii 2017], [Lounici 2019], [Ostovskii 2019] and [Jirak 2019]. We explain some of these works in more detail and establish comparisons, whenever possible, in Section 2.5.

2.3 Relative concentration inequalities for the spectrum

In this section we present the main results, Theorems 3 and 4, of the article [Araya 2020], which deal mainly with concentration inequalities of the Weyl type. We also present a Hoffman-Weilandt type inequality for the $\delta_2(\cdot,\cdot)$ as a corollary (Cor. 5).

We present a set of inequalities, one for each eigenvalue, first in a more general context, under only a summability hypothesis (see (H) below). We later specialize those inequalities under regularity hypothesis of the Sobolev type, which translate into convergence rates that put in evidence the increase in accuracy of this methodology.

We introduce the hypothesis H which assures the summability of the spectral expansion (2.1), that is

$$\|\sum_{k\geq 1}|\lambda_k|\phi_k^2\|_\infty < \infty \tag{H}$$

We have the following result under H. We define $\mathcal{V}_1(i) = \|\sum_{k=1}^{i}\phi_k^2\|_\infty$, which works as variance proxy, in the sense of concentration inequalities [Boucheron 2013], in what follows.

Theorem 3 ([Araya 2020]). *Let $W : \Omega \times \Omega \to [0,1]$ be a kernel which eigensystem satisfies H. Fix $i \in \mathbb{N}$ and define*

$$R(i) := \min\{R \in \mathbb{N} : |\lambda_i| > \sum_{k>R}|\lambda_k| \vee \sqrt{R\sum_{k>R}\lambda_k^2}\}$$

Then there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ and for $\alpha \in (0,1)$ we have

$$|\lambda_i(T_n) - \lambda_i| \lesssim |\lambda_i|\sqrt{\frac{\mathcal{V}_1(R(i))\log R(i)/\alpha}{n}}$$

with probability larger than $1-\alpha$.

The n_0 that appears in Theorem 3 depends on i in general (we will made this point more precise in the next section). One of the consequences of Theorem 3 is that $|\lambda_i(T_n) - \lambda_i|$ attains a parametric rate in terms of n, when i is fixed and n is large enough, while maintaining the scaling term $|\lambda_i|$. The latter will be fundamental to obtain sharper inequalities under more precise eigenvalue decay rate assumptions. Stated this way, this result is formally close to the CLT [Koltchinskii 2000, Cor. 5.8], which says that when the eigenvalues of K are simple (multiplicity one) then the following convergence in law holds $\lambda_i(T_n) \to \lambda_i(T_K)G_\mu(\phi_i^2)n^{-1/2}$, where G_μ is the generalized Brownian bridge associated with μ (a centered Gaussian process indexed by L^2 functions whose covariance is same as defined by μ). Notice that Theorem 3 in this respect is close to this asymptotic result, except that the variance term $\mathcal{V}(i)$ involves not only ϕ_i^2, but all the functions up to the term $R(i)$. Our $\mathcal{V}_1(i)$ is similar to those appearing in the (absolute) bounds in positive kernel literature [Shawe-Taylor 2005] and [Blanchard 2007]. In the aforementioned results, the variance term is the radius of smaller ball, in a Hilbert space, that contains the sampled feature maps. Here is the radius of the smaller ball that contains the evaluation of the eigenfunctions.

2.3.1 Regularity hypothesis

We now turn to three more specific regularity assumptions, all of which are sufficient conditions for H. We will assume that $|\lambda_i| = \mathcal{O}(f(i))$ and $\|\phi_k\|_\infty = \mathcal{O}(g(i))$, where f and g are either polynomial or exponential. More precisely, we call hypothesis H$_1$ when $|\lambda_i| = i^{-\delta}$ for some $\delta > 0$ and $\|\phi_i\|_\infty = i^s$ for $s > 0$. We will assume that $\delta > 2s+1$, which is sufficient to fulfill H. Similarly, we say that the eigensystem satisfy H$_2$ if $|\lambda_i| = e^{-i\delta}$ and $\|\phi_i\|_\infty = i^s$, with $\delta > s$. Finally, we call H$_3$ the following hypothesis $|\lambda_i| = e^{-i\delta}$ and $\|\phi_i\|_\infty = e^{is}$, with $\delta > 2s$. We will assume that $\delta, s \in \mathbb{N}$, which do not seem to be strictly necessary, but makes easier to establish a connection with the classic formulation of Sobolev regularity hypothesis. In Table 3.7.2 we summarize the assumptions we use in this section. The following theorem gives specific rates for those assumptions.

	Assumption		
H	$\|\sum_i \lvert\lambda_i\rvert\phi_i^2\|_\infty < \infty$		
	$\lvert\lambda_i\rvert$	$\|\phi_i\|_\infty$	
H$_1$	$\mathcal{O}(i^{-\delta})$	$\mathcal{O}(i^s)$	$\delta > 2s+1$
H$_2$	$\mathcal{O}(e^{-\delta i})$	$\mathcal{O}(i^s)$	$\delta > s$
H$_3$	$\mathcal{O}(e^{-\delta i})$	$\mathcal{O}(e^{si})$	$\delta > 2s$

Table 2.1: Hypotheses for the eigenvalue decay and the growth of the eigenvectors

Theorem 4. *Let W be a kernel satisfying one of the hypothesis H_1, H_2 or H_3. Then with probability larger than $1-\alpha$ we have a bound of the form*

$$|\lambda_i(T_n) - \lambda_i| \lesssim B(i,n)\log 1/\alpha$$

where $B(i,n)$ depends on the respective hypothesis and is given by the following table

Assumption	$B(i,n)$	i
H$_1(s \geq 1)$	$i^{-\delta+\frac{\delta}{\delta-1}(s+\frac{1}{2})}n^{-\frac{1}{2}}$ $i^{-\delta+1+\frac{\delta-1}{\delta}(s+\frac{1}{2})}n^{-\frac{1}{2}}$ $i^{-\delta+s+1}n^{-\frac{1}{2}}$	$1 \leq i \leq n^{\frac{\delta-1}{\delta}\frac{1}{2s+1}}$ $n^{\frac{\delta-1}{\delta}\frac{1}{2s+1}} \leq i \leq n^{\frac{1}{2s}}$ $n^{\frac{1}{2s}} \leq i \leq n$
H$_1(s = 0)$	$i^{-\delta+\frac{1}{2}}n^{-\frac{1}{2}}$	$1 \leq i \leq n$
H$_2(s \geq 1)$	$e^{-\delta i+(s+\frac{1}{2})\log i}n^{-\frac{1}{2}}$ $e^{-\delta i+s\log i}n^{-\frac{1}{2}}$	$1 \leq i \leq n^{\frac{1}{2s}}$ $n^{\frac{1}{2s}} \leq i \leq n$
H$_2(s = 0)$	$e^{-\delta i+\frac{1}{2}\log i}n^{-\frac{1}{2}}$	$1 \leq i \leq n$
H$_3(s \geq 1)$	$e^{(-\delta+s)i}n^{-\frac{1}{2}}$	$1 \leq i \leq n$

Both theorems are derived from the same set of results, but they are better adapted for different situations. The purpose of Theorem 3 is to give a rate in terms of the sample size for fixed index i. In Theorem 4, we assume a fixed sample size and allow i to vary with n. Observe that the obtained rates change depending on the value of i (relative to n). This is known phenomena in concentration inequalities, where often we have a mix between different tail regimes (frequently Gaussian and Exponential). The fact that this occurs at $i = \mathcal{O}(n^{\frac{1}{s}})$ for H_1 and H_2, when $s > 0$, and at $i = \mathcal{O}(\frac{\log n}{s})$ for H$_3$ is related to the transition between $\sqrt{\frac{\mathcal{V}_1(R)}{n}}$ and $\frac{\mathcal{V}_1(R)}{n}$, which appears frequently in concentration inequalities. This will be clear from the explanation of our approach in Section 2.4. In Theorem 3 this is less important as the focus is when n is large, while i is fixed, and the term $\mathcal{O}(1/\sqrt{n})$ prevails. Observe that Theorem 4 gives rates that are better than parametric under H_1 and exponential in cases H_2 and H_3. This in line with some recent results such as [Belkin 2018, Thm.2]. A more detailed comparison with previous results in the literature is postponed to Section 2.5. In the next section we explain the ideas behind the proof of the main results.

We end this section with a corollary that shows that the previous results can be used to control the deviation of more than one eigenvalue at the time. We prove the following Hoffman-Weidlandt style bound for the $\delta_2(\cdot,\cdot)$ metric

Corollary 5. *For a kernel K satisfying H_2 or H_3, then we have with probability larger than $1-\alpha$*

$$\delta_2(\lambda(T_n),\lambda(T_K)) = \mathcal{O}_\alpha\Big(\frac{1}{\sqrt{n}}\Big)$$

If K satisfy H_1 with the additional assumption that $\delta > 2s+2$, the same conclusion holds.

While this result might be deduced from the absolute bounds for positive operators (by a simple decomposition in positive and negative parts[3]), presented for example in [Rosasco 2010]. This method still offers advantages. For instance, it allows for the consideration of portions of the spectrum other than the full spectrum or a single eigenvalue. This could be useful for algorithms using a group of eigenvalues for discovering a low dimensional structure in the data, such as kernel PCA, or for the algorithm (HEiC) for recover latent distances in graphs that will be described in Chapter 3.

2.4 Three step method: approximation, perturbation and concentration.

The purpose of this section is to explain the three step approach used in the proof of Theorems 3 and 4, as developed in [Araya 2020]. Each step in this approach is named after the family of techniques in display. In the approximation step we use K_R, a rank R approximation of K to find a convenient factorization for T_n. This gives a framework where T_n can be seen as a random perturbation of T_K. In the perturbation step, we use deterministic matrix perturbation inequalities to quantify the deviation between $\lambda(T_K)$ and $\lambda(T_n)$. At the end of this step we obtain a scaling bound that depend on random error terms (which are written as the operator norm of two random matrices that appear in the perturbation). Finally, in the concentration step we use concentration inequalities for U-statistics to control the error terms. We now describe each step in more detail.

Approximation step: The approximation step builds upon a truncation approach, which is also used for example in [Koltchinskii 2000], [De Castro 2020] and [Braun 2006], where we fix $R \in \mathbb{N}$ and decompose W into two terms:

[3] Note that some technical conditions are required for this to work. Under our hypothesis H they are guaranteed.

$$(T_n)_{ij} = \frac{1}{n} W_R(X_i, X_j) + \frac{1}{n}(W - W_R)(X_i, X_j)$$

We call the first term of the right hand side, the *R-truncated kernel matrix.* Define the *residual matrix* (the error from the approximation) as

$$(E_R)_{ij} := \frac{1}{n}(W - W_R)(X_i, X_j) = \sum_{k>R} \lambda_k \phi_k(X_i)\phi_k(X_j)$$

where the second equality is justified by the assumption H, which implies the pointwise equality. The R-truncated kernel matrix can be written as a multiplicative perturbation of a diagonal matrix. More specifically, we have the following factorization

$$\frac{1}{n} W_R(X_i, X_j) = \Phi_R \Lambda_R \Phi_R^T$$

where Φ_R is the $n \times R$ matrix with columns $1/\sqrt{n}(\phi_k(X_1), \phi_k(X_2), \cdots, \phi_k(X_n))^T$ and Λ_R is a diagonal matrix with $\lambda_1, \lambda_2, \cdots, \lambda_R$ in the diagonal. Thus, the normalized kernel matrix can be written as an additive perturbation of the R-truncated matrix by the residual matrix

$$T_n = \Phi_R \Lambda_R \Phi_R^T + E_R \tag{2.2}$$

From the previous equality is already possible to obtain scaling bounds, as it is done in [Braun 2006]. The idea is that the first term has a structure that makes it compatible with standard tools of matrix concentration (after using a deterministic multiplicative perturbation theorem). The residual term has less structure, but its operator norm will be small in comparison to the R-truncated matrix, provided that R is well chosen. In [Braun 2006] and [De Castro 2020] the classic (absolute) Weyl inequality is used. Intuitively speaking, the problem with that approach is that in absolute perturbation we consider the effect of the residual term over all the eigenvalues of the truncated matrix is uniform, which should not be the case, in light of the asymptotic results in [Koltchinskii 2000]. To overcome this, we introduce another factorization instead, which allows to better exploit the multiplicative perturbation framework. The factorization is as follows

$$T_n = (\Phi_R|\Phi_R^{\perp}) M (\Phi_R|\Phi_R^{\perp})^T + A \tag{2.3}$$

where

$$M = \begin{pmatrix} \Lambda_R & 0 \\ 0 & M_{>R} \end{pmatrix}$$

the columns of matrix $\Phi_R^{\perp}$ are an orthonormal basis of the orthogonal complement to the space spanned by the columns of Φ_R. Assume for the moment that the columns of Φ_R are linearly independent. On that event, define the projection matrices $P_1 := \Phi_R(\Phi_R^T\Phi_R)^{-1}\Phi_R^T$ and $P_2 := \Phi_R^{\perp}\Phi_R^{\perp T}$, the matrices $M_{>R}$ and A are specified by

$$\begin{aligned} M_{>R} &:= \Phi_R^{\perp T} E_R \Phi_R^{\perp} \\ A &:= P_1 E_R P_2 + P_2 E_R P_1 + P_1 E_R P_1 \end{aligned}$$

Notice that the columns of $\Phi_R^{\perp}$ are orthonormal, which is not the case of the columns of Φ_R. In words, we decompose the residual matrix according to its projection onto the space generated by the columns of Φ_R and its orthogonal complement.
Perturbation step: Define $\tilde{M} := (\Phi_R|\Phi_R^{\perp})M(\Phi_R|\Phi_R^{\perp})^T$. We first use Weyl's perturbation theorem to obtain the following

$$|\lambda_i(T_n) - \lambda_i(\tilde{M})| \leq \|A\|_{op} \tag{2.4}$$

Now we use a relative multiplicative perturbation theorem known as Ostrowskii's inequality, in the non-square version given by Cor. 17 (this inequality is stated in the case of non decreasing eigenvalues, but it is still valid for the ordering we use here, see Remark 1) to obtain for all $1 \leq i \leq n$

$$|\lambda_i(\tilde{M}) - \lambda_i(M)| \leq |\lambda_i(M)| \|(\Phi_R|\Phi_R^{\perp})^T(\Phi_R|\Phi_R^{\perp}) - \mathrm{Id}_n\|_{op} = |\lambda_i(M)| \|\Phi_R^T\Phi_R - \mathrm{Id}_R\|_{op} \tag{2.5}$$

where the last equality comes from the fact that $\Phi_R^{\perp}$ has orthonormal columns. Using (2.5) and (2.4) we obtain for all $1 \leq i \leq n$

$$|\lambda_i(T_n) - \lambda_i(M)| \leq |\lambda_i| \|\Phi_R^T\Phi_R - \mathrm{Id}_R\|_{op} + \|A\|_{op} \tag{2.6}$$

Because of the block structure of M, we have that $\lambda(M) = \lambda(\Lambda_R) \cup \lambda(M_{>R})$. The eigenvalues $\lambda(\Lambda_R)$ are deterministic, while $\lambda(M_{>R})$ is a random set. By the definition of $M_{>R}$, the following trivial bound holds

$$\lambda_i(M_{>R})| \leq \|\Phi_R^{\perp T} E_R \Phi_R^{\perp}\|_{op}, \text{ for all } 1 \leq i \leq n-R$$

Concentration step: We use concentration inequalities to control with high probability the terms in the right hand side of (2.6) and they will also serve us to characterize the random set $\lambda(M_{>R})$. As we already mention, inequalities for quantities of the form $\|\Phi_R^T\Phi_R - \mathrm{Id}_R\|_{op}$ are well stablished. For instance, given that $\Phi_R^T\Phi_R$ can be written as a sum of independent random matrices, we can use some version of the non-commutative Bernstein inequality. Using, for example, the matrix Bernstein inequality (see Thm. 19) we obtain

Proposition 6. *With probability larger than $1-\alpha$ we have*

$$\|\Phi_R\Phi_R^T - \mathrm{Id}_R\|_{op} \lesssim \frac{\mathcal{V}_1(R)\log R/\alpha}{n} \vee \sqrt{\frac{\mathcal{V}_1(R)\log R/\alpha}{n}}$$

On the other hand, for the terms $\|(\Phi_R^T\Phi_R)^{-1}\Phi_R^T E_R \Phi_R(\Phi_R^T\Phi_R)^{-1}\|_{op}$ and $\|\Phi_R^{\perp} E_R \Phi_R^{\perp T}\|_{op}$, the standard tools in matrix concentration inequalities, do not seem to apply as smoothly. For instance, when E_R has infinite rank, it cannot be expressed as a finite sum of independent matrices. Some concentration inequalities, such a the matrix bounded differences inequality [Mckey 2016, Cor. 6.1] or others obtained by the matrix Stein method [Mckey 2014] can be applied in this case, but they demand strong conditions such as almost sure control of a matrix variance proxy in the semi-definite order. In addition, in this case they deliver suboptimal results. We opt for using a rougher matrix norm inequality, for example using the Frobenius norm to control the operator norm, and we then use concentration inequalities for U-statistics, such as those in [Gine 2000, Thm.3.3] or [Houdré 2003, Thm.3.4]. This is can be seen as rough application of the *comparison method* described in [Van Handel 2017], which consists in find an easier-to-bound random process majorizing the operator norm. Finding a tight majorizing random process is, in general, a challenging task and no canonical way to do this is known, to the best of our knowledge. The fact that we use a rougher bound for the matrix norm will be compensated by optimizing the choice of R, which helps reducing the impact of this inaccuracy.

We have the following proposition which gives a tail bound for the terms $\|E_R\phi\|_{op}$, for $\phi \in \{\phi_1, \cdots, \phi_R\}$, and $\|{\Phi_R^{\perp}}^T E_R \Phi_R^{\perp}\|_{op}$.

Proposition 7. *We have with probability larger than* $1-\alpha$

$$\sqrt{\sum_{l=1}^{R} \|E\phi_l\|_{op}^2} \lesssim_\alpha \sqrt{\frac{1}{n} b_{2,R} \mathcal{V}_1'(R)} =: \gamma_1(n,R) \tag{2.7}$$

$$\|{\Phi_R^{\perp}}^T E_R \Phi_R^{\perp}\|_{op} \lesssim_\alpha b_R + \sqrt{\frac{\mathcal{V}_2(R) b_R}{n}} \vee \frac{\mathcal{V}_2(R)}{n} =: \gamma_2(n,R) \tag{2.8}$$

where

$$b_R := \sum_{k>R} |\lambda_k|, \quad b_{2,R} := \sum_{k>R} \lambda_k^2, \quad \mathcal{V}_1'(R) := \sum_{k=1}^{R} \|\phi_k\|_\infty^2, \quad \mathcal{V}_2(R) := \|\sum_{k>R} \lambda_k \phi_k \otimes \phi_k\|_\infty$$

For $\tau > 0$, we define the event

$$\mathcal{E}_\tau := \{\omega \in \Omega \text{ s.t } \|\Phi_R^T \Phi_R(\omega) - \mathrm{Id}_R\|_{op} < \tau\}$$

Define $\tau_{n,R,\alpha} := \sqrt{\frac{\mathcal{V}_1(R) \log R/\alpha}{n}}$. The following lemma, which proves that the event $\mathcal{E}_{\tau_{n,R,\alpha}}$ holds with high probability, is proven using Proposition 6.

Lemma 8. *For* $R < n$ *we have*

$$\mathbb{P}\Big(\mathcal{E}_{\tau_{n,R,\alpha}}\Big) \geq 1-\alpha$$

Lemma 9. *Let* $\mathrm{Sp}(\Phi_R)$ *be the linear span of* $\phi_1, \cdots, \phi_R$*. It holds*

$$\|A\|_{op} \lesssim \max_{\phi \in \mathrm{Sp}(\Phi_R), \|\phi\|=1} \|E\phi\|$$

Let $\mathcal{E}_\alpha$ *be the event such that* (2.7) *holds. In the event* $\mathcal{E}_\alpha \cap \mathcal{E}_{\tau_{n,R,\alpha}}$ *we have*

$$\|A\|_{op} \lesssim_\alpha \frac{1}{1-\tau_{n,R,\alpha}} \gamma_1(n,R)$$

For $\alpha \in (0,1)$ and $R < n$ we have, using Lemma 9, Lemma 8 and Prop. 7

$$\|A\|_{op} \lesssim_{\alpha} \frac{1}{1-\tau_{n,R,\alpha}} \gamma_1(n,R) \tag{2.9}$$

with probability larger than $1 - 2\alpha$. The following two propositions will allow us to control $|\lambda_i(T_n) - \lambda_i|$ for a fixed $R \in \mathbb{N}$.

Proposition 10. *Assume that $R \in \mathbb{N}$ is such that $\tau_{n,R,\alpha} < 1$. Then with probability larger than $1-\alpha$ we have, for $i < R$*

$$|\lambda_i(T_n) - \lambda_i| \lesssim_{\alpha,\tau} \big(|\lambda_i| \vee \gamma_2(n,R)\big)\sqrt{\frac{\mathcal{V}_1(R)\log R}{n}} + \gamma_1(n,R) \tag{2.10}$$

Proposition 11. *Fix $R \in \mathbb{N}$. We have, with probability larger than $1-\alpha$ for $i > R$*

$$|\lambda_i(T_n) - \lambda_i| \lesssim_{\alpha} \gamma_2(n,R)$$

The proof of Theorem 3 uses Proposition 10. Indeed, it is easy to see that $\gamma_2(n,R) \to 0$ when $R \to \infty$ and the same is true for $\gamma_1(n,R)$(actually this terms also converge to 0 if $n \to \infty$). For a fixed i and for n large enough, we can always choose R to satisfy $\lambda_i > \gamma_2(n,R)$ and $\lambda_i > \gamma_1(n,R)$, then the Proposition 10 will imply the bound in Theorem 3. For Theorem 4, on the other hand, we use either Prop. 10 or Prop. 11 depending on the relative position of i with respect to n. The fact that we have explicit assumptions on the eigenvalues and eigenfunctions allow us to make the relation between λ_i, $\gamma_1(n,R)$ and $\gamma_2(n,R)$ more precise. We include a sketch of the proof in Section 2.11.

Remark 1 (Ostrowski's for the decreasing in absolute value ordering)**.** *As we mentioned above, the Ostrowski's inequality is formulated in the case of non-decreasing(or equivalently non-increasing) ordering. Nonetheless, it is still valid for the decreasing ordering in the absolute value. Indeed, if $\{\lambda_{\sigma(i)}\}_{1\le i\le n}$ is an ordering of the eigenvalues, that is $\sigma : [n] \to [n]$ is bijective, we can reorder them in the non increasing order by applying a transformation $\sigma^{\uparrow}$ to each $\sigma(i)$, then apply the Owstroski's inequality and finally apply ${\sigma^{\uparrow}}^{-1}$. The key here is that this reordering process is applied to a finite matrix, because some orderings are not compatible with the operator T_W full spectrum (the increasing ordering cannot be applied to the spectrum of an indefinite operator, given that 0 is an accumulation point).*

2.5 Asymptotic rate analysis

The rates obtained in Theorem 4 are expressed in terms of i and n, which is natural when we are interested in a fixed i, or purely in terms of n (by replacing $i = n^{\log i/\log n}$). Under H_1, we obtained a parametric rate in terms of n(concentration term) and $i^{-\delta+(s+1/2)g(i)}$(scaling and variance term), where $g(i) \le 2$. This in line with the CLT in [Koltchinskii 2000], where the same scaling and concentration terms appears. In that sense, the concentration and scaling terms seems optimal, while there might be room for some improvement in the variance term. If we allow i to vary with n, we obtain rates that are fully expressed in terms of n, in which case they are always faster than $\mathcal{O}(n^{-1/2})$ for all three hypothesis H$_1$, H$_2$ and H$_3$. This implies that our result are more accurate than all absolute type bounds, as those obtained in [Shawe-Taylor 2005], [Blanchard 2007], [Rosasco 2010], for example, which all give $\mathcal{O}(\frac{1}{\sqrt{n}})$ bound.

In [Braun 2006] the authors obtain scaling bounds, in the p.s.d case, which are slower than those given in Theorem 4 under the three hypothesis H$_1$, H$_2$ and H$_3$. They do not assume explicit growth rates, but formulate their result under the assumption of bounded eigenfunctions (which fall in our framework with $s = 0$) and bounded kernel function(which is implied by H). For example, in the case of polynomial decay of the eigenvalues and bounded kernel, they obtain an error rate of $\mathcal{O}(n^{\frac{1-\delta}{2\delta}}\sqrt{\log n})$, which is slower than $\mathcal{O}(n^{-1/2})$ in terms of n. We observe that for i fixed we obtain a better rate for the error in terms of n. Indeed, the rate in Theorem 4 under H$_1$ is $\mathcal{O}(n^{-\frac{\log i}{\log n}(\delta-1)-1/2)})$, which is faster than $\mathcal{O}(n^{-\frac{1}{2}})$ provided that $\delta > 1$ (which we have to assume in order to satisfy H). Similar comparison can be stablished in the case of exponentially decay eigenvalues, where they obtain a parametric rate(except for logarithmic terms) and ours is exponential. Also, in the case of fixed i as in Theorem 3, we avoid the cumbersome bias terms present for example in [Braun 2006, Thm. 3].

In [Lounici 2019] and [Ostovskii 2019] similar relative bounds are proven, which are in line with sample covariance concentration of [Koltchinskii 2017]. Their bounds for the difference of the empirical eigenvalue $\hat{\lambda}_i$ and the population eigenvalue λ_i of covariance matrices is of the form $|\hat{\lambda}_i - \lambda_i| \lesssim \lambda_i\sqrt{\frac{r(S)}{n}}$, where $r(S) = Tr(S)/\|S\|_{op}$ and S is the population covariance matrix. In the case of [Ostovskii 2019], a different variance term is introduced, which depend on a regularization step based on shifting up the eigenvalues of S. Those results are formally similar to those in Theorem 3, but the variance term differ. Observe that, since $\|S\|_{op}$ acts as a normalization term, their $r(S)$ is a constant(do not change with i). While is true that the term $r(S)$ can be smaller than $\mathcal{V}_1(R(i))$, note that the first constitute a uniform control over all indices. Moreover, their work is devoted to the more particular case of sample covariance estimation, where one of the main assumptions is that the random vectors, of which S is the population covariance, have a fixed subgaussian norm which do not change in terms of the ambient dimension. This is not the case in our context, as explained in Section 2.4 above, given that we consider and abstract space Ω and we do not have a fixed Euclidean space where the columns of Φ_R belong to, but we rather define $\mathbb{R}^R$ given the truncation parameter, which is later optimized. To manage the increase in variance carried by the term $\mathcal{V}_1(\cdot)$ is one of the main technical difficulties we tackle in [Araya 2020]. In addition, the works [Lounici 2019] and [Ostovskii 2019] are formulated in the positive case approach and extensions to the indefinite case are not discussed.

A different approach is used in [Belkin 2018]. Using approximation theoretic methods, they obtain a measure independent result, from which a rate for the empirical measure can be easily deduced. Indeed, the rate obtained in [Belkin 2018, Thm.2] for $|\lambda_i(T_n) - \lambda_i(T_W)|$ is $\mathcal{O}(e^{-ci^{1/d}})$, where c is a positive constant and W is a positive, radially symmetric, infinitely differentiable kernel defined on $\mathbb{R}^d$. This represents an intermediate regime between H_1 and H_2. Their rate do not seem to depend explicitly on the eigenfunctions growth, however the fact that the kernel is highly regular and radially symmetric would have an effect (this shares similarities with the case of dot product kernels presented in Section 2.7 below). At least formally, our results are aligned with those in [Belkin 2018], in the sense that when an exponential rate of the eigenvalues of T_W is observed, the deviation $|\lambda_i(T_n) - \lambda_i(T_W)|$ will have an exponential rate (the scaling term prevails over the concentration term). It is worth mentioning that the approximation theoretic methods used in [Belkin 2018] rely heavily on the RKHS technology and do not extend automatically to non positive case. In addition, their result do not consider lower regularity kernels which are very common in the network analysis, for instance (see Section 2.7 and Chapter 3). Extension of this approach to the indefinite case, relaxing the symmetric and high regularity hypothesis, might be possible using for example the Krein spaces framework. That constitutes a substantially different approach, which we leave for future work.

	Assumption	Rate
Thm. 4	$\mathrm{H}_1(s=0)$	$n^{-\frac{\log i}{\log n}(1-\delta)-1/2}$
Thm. 4	H_3	$e^{-\delta i+(2s+1)\frac{\log i}{\log n}}n^{-1/2}$
Thm. 3	H	$\lvert\lambda_i\rvert\sqrt{\frac{\mathcal{V}_1(i+m_i)}{n}}$
[Braun 2006, Thm.3]	$\mathrm{H}_1(s=0)$	$n^{\frac{1-\delta}{2\delta}}$
[Belkin 2018]	H_2[4]	$e^{-i^{1/d}}$
[Lounici 2019],[Ostovskii 2019]	H[5]	$\lvert\lambda_i\rvert\sqrt{\frac{r(S)}{n}}$

Table 2.2: Rate comparisons with related relative Weyl-type inequalities in the literature, we omit logarithmic terms.

2.6 Classical Sobolev regularity conditions

The fact that a given kernel W satisfy any of the hypothesis H_1, H_2 and H_3, of Theorem 4, is not necessarily easy to verify. If an eigenfunctions basis is known, such as the case of spherically symmetric kernels treated in Section 2.7 below, the eigenvalues can be obtained by the computing the integral of the product of the kernel with the eigenfunctions of the basis. There is no guarantee that an analytic close solution exist in general, but in practice this procedure can be done numerically. On the other hand, when the eigenfunctions of the kernel are not known, we are left to solve often complicated differential equations. For that reason is useful to have an equivalent notion of regularity at hand.

The decrease in the eigenvalues appears naturally as regularity hypothesis of the Sobolev-type. Indeed, given a measurable metric space $(\mathcal{X}, \kappa, \nu)$ where κ is a distance and ν a probability measure, we suppose that $\{\varphi_k\}_{k\in\mathcal{J}}$ is an orthonormal basis of $L^2(\mathcal{X}, \nu)$, where $\mathcal{J}$ is a countable set. We define the weighted Sobolev space S_ω with associated positive weights $\omega = \{\omega_j\}_{j\in\mathcal{J}}$ as

$$S_\omega(\mathcal{X}) := \Big\{ f \overset{L^2}{=} \sum_{k\in\mathcal{J}} \hat{f}(k)\varphi_k \text{ s.t } \|f\|_\omega^2 := \sum_{k\in\mathcal{J}} \frac{|\hat{f}(k)|^2}{\omega_k} < \infty \Big\}$$

Take the measurable metric space (Ω, ρ, μ) and consider $\mathcal{X} = \Omega^2$ and $\nu = \mu \times \mu$. If ϕ_k is a basis of $L^2(\Omega, \mu)$ then a basis for $L^2(\mathcal{X}, \mu)$ is given by $\{\varphi_k\}_{k,l}$ where $\varphi_{k,l} = \phi_k \otimes \phi_l$. Observe that here $\mathcal{J} = \{(k,l)\}_{k,l\in\mathbb{N}}$. We note that for a kernel W in $S_\omega(\mathcal{X})$ with eigenvalues λ_k and eigenvectors ϕ_k, we have $\hat{f}(k,l) = \lambda_k \delta_{kl}$. If we want that the series in definition of the Sobolev space to converge, it is sufficient that $\lambda_k^2 \frac{1}{\omega_k} = \left(\frac{1}{k^{1+\delta'}}\right)$ where $\delta' > 0$. This allow to control the decay behavior of λ_k by direct comparison to ω_k. When Ω is an open subset of $\mathbb{R}^d$, the classical definition of weighted Sobolev spaces makes use of the (weak)-derivatives of a function. If $\varrho : \Omega \to [0, \infty)$ is a locally integrable function, we define the weighted Sobolev space $\mathcal{W}_2^p(\Omega, \varrho)$ as the normed space of locally integrable functions $f : \Omega \to \mathbb{R}$ with p weak derivatives such as the following norm is finite

$$\|f\|_{p,\varrho} = (\int_\Omega |f(x)|^2 d\varrho(x))^{\frac{1}{2}} + (\sum_{|\alpha|=p} |D^\alpha f(x)|^2 d\varrho(x))^{\frac{1}{2}}$$

where α is a multiindex and D^α are the weak derivatives.

For a symmetric kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ we can define the Sobolev regularity by the canonical embedding of $\mathbb{R}^d \times \mathbb{R}^d$ into $\mathbb{R}^{2d}$, but it seems more natural (see [Xu 2017, sect. 2.2]) to say that the kernel satisfies the weighted Sobolev condition if $K(\cdot, x) \in \mathcal{W}_2^p(\Omega, \varrho)$ for all $x \in \Omega$. However, in some cases as in the *dot product kernels*, where there exists a real function $f : \mathbb{R} \to [0,1]$ such as $K(x,y) = f(\langle x, y\rangle)$, it is even more natural to say that K that satisfies the Sobolev condition with weight $\varrho : \mathbb{R} \to \mathbb{R}$ if $f \in \mathcal{W}_2^p(\mathbb{R}, \varrho)$. Intuitively speaking, given that f is defined on $\mathbb{R}$, it seems natural to carry out the analysis in one dimension.

In [Nicaise 2000] is proved that in the one dimensional case, both definitions of weighted Sobolev spaces are coincident. Otherwise stated, the following equality between metric spaces holds $S_w([-1,1]) = \mathcal{W}_2^2([-1,1], \varrho_\gamma)$ where $\omega = \{\omega_k\}_{k\in\mathbb{N}} = \frac{1}{1+\nu_k}$, with $\nu_k = k(k+d-1)$ and $\varrho_\gamma(x) = (1-x^2)^{\frac{d-3}{2}}$. Here we recognize in ν_k the sequence of eigenvalues of the Laplace-Beltrami operator on $\mathbb{S}^{d-1}$ and ϱ_γ is the weight that defines the orthogonality relations between the Gegenbauer polynomials $G_l^\gamma(\cdot)$ with $\gamma = \frac{d-2}{2}$. That means that two Gegenbauer polynomials of different degrees G_k^γ, G_l^γ with $k \neq l$ are orthogonal in $L^2([-1,1], \varrho_\gamma)$, which is the space of square integrable functions defined in $[-1,1]$ with the weight ϱ_γ. We denote $\|\cdot\|_{2,\gamma}$

[4] Here is a result more than an assumption. The actual assumptions are the high regularity of the kernel and its radial symmetry.

[5] The hypothesis here are of the type on the subgaussian norm of random vectors.

the norm in $L^2([-1,1],\varrho_\gamma)$, that is $\|f\|_{2,\gamma}^2 = \int f^2(t)\varrho_\gamma(t)dt$. In the next section, we explore this case in more detail and highlight the connection with random geometric graphs.

2.7 Dot product kernels

In this section we will consider the space $\Omega = \mathbb{S}^{d-1}$, with $d \geq 3$, equipped with ρ the geodesic distance and the measure σ, which is the surface (or uniform) measure normalized to be a probability measure. Let $f : [-1,1] \to [0,1]$ be a measurable function of the form $K(x,y) = f(\cos\rho(x,y))$. Note that the geodesic distance on the sphere is codified by the inner product, that is $\rho(x,y) = \arccos\langle x,y\rangle$. Thus we directly assume, here and thereafter, that W only depends on the inner product, that is

$$K(x,y) = f(\langle x,y\rangle)$$

This family of kernels are usually known as *dot product kernels* and they are rotation invariant, that is $K(x,y) = K(Ax,Ay)$ for any rotation matrix A, and its associated integral operator T_K is a convolution operator. Similar to the context of Fourier analysis of one dimensional periodic functions, in this case we have a fixed Hilbertian basis of eigenvectors that only depends on the space Ω, but not on the particular choice of kernel K. The aforementioned basis is composed by the well-known *spherical harmonics* [Dai 2013, chap. 1], which play the role of the Fourier basis in this case. For each $l \in \mathbb{N}$ we have an associated eigenspace $\mathcal{Y}_l$, known as the space of spherical harmonics of order l. Let $\{Y_{jl}\}_{j=1}^{d_l}$ be an orthonormal basis of $\mathcal{Y}_l$ and define $d_l = \dim(\mathcal{Y}_l)$, then by [Dai 2013, cor. 1.1.4]

$$d_l = \binom{l+d-1}{l} - \binom{l+d-3}{l-2} = \mathcal{O}(l^{d-2}) \tag{2.11}$$

for $l \geq 2$ and $d_0 = 1$,$d_1 = d$. The second equality follows easily from the definition. We define λ_l^*, the eigenvalue of T_K associated with the corresponding space $\mathcal{Y}_l$. We use the $*$ subcript to difference this indexation(who follows the spherical harmonics order) from the decreasing order indexation $\{\lambda_i\}_{i\geq 1}$. As sets $\{\lambda_i^*\}_{i\geq 0}$ and $\{\lambda_i\}_{i\geq 1}$ are equal (have the same elements), but in $\{\lambda_i^*\}$ the eigenvalues are counted without multiplicity (except if λ is associated to more than one $\mathcal{Y}_l$ [6]). This seems more natural in this case, but have to keep this in mind when applying Theorems 3 and 4. In this setting, the expansion (2.1) becomes

$$f(\langle x,y\rangle) = \sum_{l\geq 0} \lambda_l^* \sum_{j=0}^{d_l} Y_{jl}(x)Y_{jl}(y) \tag{2.12}$$

[6] In which case appears repeated a number of times equals to the number of $\mathcal{Y}_l$'s to which is associated.

On the other hand, the *Addition Theorem* for spherical harmonics [Dai 2013, eq. 1.2.8] gives

$$Z_l(x,y) = \sum_{j=0}^{d_l} Y_{jl}(x) Y_{jl}(y) \tag{2.13}$$

and the preceding equality does not depend on the particular choice of basis $\{Y_{jl}\}_{j=1}^{d_l}$. The Z_l are called the *zonal harmonics.* So based on (2.12) we have the following

$$f(\langle x,y\rangle) = \sum_{l\geq 0} \lambda_l^* Z_l(x,y) \tag{2.14}$$

An important property is that each zonal harmonic $Z_l(x,y)$ is a multiple of the Gegenbauer (ultraspherical) polynomial of level l, hence it only depends on the inner product of $x, y \in \mathbb{S}^{d-1}$. The following classic result in Harmonic Analysis [Dai 2013, Thm.1.2.6, Cor. 1.2.7] makes the previous statement more precise

Proposition 12. *For any $x, y \in \mathbb{S}^{d-1}$, $l \in \mathbb{N}$, $d \geq 3$ and $\gamma = \frac{d-2}{2}$*

$$Z_l(x,y) = c_l G_l^\gamma(\langle x,y\rangle) = c_l \sqrt{d_l}\tilde{G}_l^\gamma(\langle x,y\rangle)$$

where $c_l := \frac{l+\gamma}{\gamma} = \frac{2l+d-2}{d-2}$, G_l^γ is the l-th Gegenbauer (ultraspherical) polynomial and $\tilde{G}_l^\gamma = G_l^\gamma/\|G_l^\gamma\|_{2,\gamma}$. Furthermore, for any $l \in \mathbb{N}$, Z_l attains its maximum in the diagonal, that is

$$\max_{x,y\in\mathbb{S}^{d-1}} |Z_l(x,y)| = |Z_l(x,x)| = d_l$$

Remark 2 (Eigenvalue computation)**.** *From Proposition 12 we derive a simple formula to compute the eigenvalues, using the orthogonality relations between Gegenbauer polynomials. We recall that given $\varrho_\gamma(x) = (1-x)^\gamma$ (the Sobolev weight defined in Section 2.6) we have*

$$\int_{\S^d} \tilde{G}_k^\gamma(t)\tilde{G}_l^\gamma(t)\varrho_\gamma(t)dt = \delta_{kl}$$

Defining $b_d = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}$ we have

$$\lambda_l^* = \left(\frac{c_l b_d}{d_l}\right)\int_{-1}^{1} f(t)G_l^\gamma(t)\varrho_\gamma(t)dt = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}\frac{l!}{(2d-2)^{(l)}}\int_{-1}^{1} f(t)G_l^\gamma(t)\varrho_\gamma(t)dt \tag{2.15}$$

where $(a)^{(i)} = a\cdot(a+1)\cdots(a+i-1)$ is the rising factorial or (rising) Pochammer symbol.

What precedes means that, in this framework, the growth rate of the eigenvector is known and fixed, and the fulfillment of the hypotheses of Theorems 3 and 4 depends on the eigenvalue decay rate only, which can be verified using formula (2.15) above. This is one of the main reasons on why our results are appealing in this context.

Given Proposition 12, the hypothesis H is implied by the following

$$\sum_{l\geq 0} |\lambda_l^*| d_l < \infty$$

Because of the explicit value of d_l (given in (2.11)), we get that $|\lambda_l^*| = O(l^{1-d-\varepsilon})$, for any $\varepsilon > 0$, it is sufficient for H to hold. The following lemma is a consequence of the Addition Theorem eq.(2.13)

Lemma 13. *For any $i \in \mathbb{N}$ we have that*

$$\mathcal{V}_1(i) = \mathcal{O}(i)$$

Consequently, for any W such that $\lambda_i = O(i^{-\delta})$ with $\delta > 1$ hypothesis H_1 is satisfied. If $\lambda_i = O(e^{-\delta i})$ with $\delta > 0$, then hypotheses H_2 and H_3 are satisfied.

Remark 3. *Observe that in the previous lemma, the eigenvectors growth do not depend on d in the indexation induced by $\{\lambda_i\}_{i\geq 1}$. However, given that the eigenvalues satisfy $\sum_{l\geq 0} |\lambda_l^*| d_l \leq \infty$ and that $d_l = \mathcal{O}(l^{d-1})$, the dimension has a direct effect on hypotheses H_1 and H_2. Indeed, for a kernel on space $\mathbb{R}^d$ with high d, the hypothesis $|\lambda_i| = \mathcal{O}(i^{-\delta})$, for any δ, will be more restrictive compared to a kernel defined on a lower dimensional space, simply because the multiplicity constraint.*

The following lemma allow us to relax the hypotheses of Thm.4 and to obtain sharper results in this case.

Lemma 14. *Let W be a kernel such that $\mathcal{V}_1(i) = \mathcal{O}(i)$ for all i and $\mathcal{V}_2(R) = \mathcal{O}(\sum_{i>R} |\lambda_i|)$, then the results of Theorem 4 are valid with $s = 0$.*

Corollary 15 ([Araya 2020]). *Let $W(x,y) = f(\langle x,y\rangle)$ be a dot product kernel in $\mathbb{S}^{d-1}$. Suppose that f is in the Sobolev space $\mathcal{W}_2^p([-1,1],\varrho')$ where $\varrho'(t) = (1-t^2)^{\frac{d-3}{2}}$.Then there exists $\varepsilon > 0$ such that for $\alpha \in (0,1)$ we have with probability larger than $1-\alpha$, for $1 \leq i \leq n$*

$$|\lambda_i(T_n) - \lambda_i| \lesssim_\alpha i^{-\delta+1/2} n^{-1/2} \tag{2.16}$$

with $\delta = \frac{p+\varepsilon}{d-1} + \frac{1}{2}$.

Remark 4. *Observe a similar framework to the one presented in this section was studied in [De Castro 2020] in the context of graphon estimation through the spectra. They bound the $\delta_2(\cdot,\cdot)$ metric, which implies using Weyl's inequalities a rate $\mathcal{O}_\alpha(n^{-\frac{\delta}{2\delta+d-1}})$ for $|\lambda_i(T_n) - \lambda_i|$, which is slower than the rate in Corollary (2.16).*

2.7.1 Connection with Random Geometric Graphs

Dot product kernels are related with the model of Random Geometric Graphs (RGG), in the dense case, via the graphon formalism [Borgs 2012]. The RGG has found many application in the fields of wireless networks [Franceschetti 2008], biology [Higham 2008b] and physics [Cunningham 2017]. Also it has theoretical importance, since often a graph in this class (proximity graphs for instance) are constructed in the first steps of clustering or embedding algorithms. A thorough study from the probabilistic point of view can be found in [Penrose 2003] and some interesting problems in the field of statistical analysis of networks is described [Bubeck 2017].

To construct a graph from a kernel, we use the W-random graph model, described in [Lovasz 2012, Sec.1] and in Chapter 1. This require an additional sampling step, in comparison with the results presented here. We can think this as having two sources of randomness. Indeed, we have $\{X_i\}_{1\leq i\leq n} \in \Omega$, where Ω is our "geometric" space (this will be further developed in the next chapter) and we construct the kernel matrix. In the context of graphs, the kernel matrix is often called *probability matrix* [Klopp 2017a] from which the adjacency matrix A of a graph is obtained by sampling independent (except for symmetry constraints) Bernoulli's random variables. More specifically, the entries $\{A_{ij}\}_{1\leq i<j\leq n}$ are independent and

$$\mathbb{P}(A_{ij}=1)=K(X_i,X_j)$$

In the case of dot product kernels on the sphere, the spectral approach has found applications in testing for geometry [Bubeck 2016], where the authors construct a test based on a triangle statistic to decide if a graph has geometric structure(the alternative hypothesis of the test is that it comes from a RGG model). The null hypothesis is that the graphs comes from a Erdös-Rényi model (geometrically structureless). Note that the quantity of triangles can expressed in terms of the spectra of the graph as a constant times $\sum_{i=1}^{n}\lambda_i^3(A)$, where A is the adjacency matrix. So essentially, [Bubeck 2016] presents a spectral test. Given that concentration of the triangle statistics plays a big role in that test, we believe that our results may find applications in testing problems for geometric graphs and for other models as well.

The spectral point of view for RGG is also adopted in [De Castro 2020], where the authors study the problem of graphon estimation in the case of angular RGG model. The authors prove that the graphon function can be reconstructed using the Gegenbauer basis and an estimation of each eigenvalue. Certainly a better knowledge of the eigenvalue fluctuation, using Cor. 5 for instance, will improve upon the results presented there.

Finally, we found applications in the problem of estimating the latent distances in the RGG model. In [Araya 2019] we propose an spectral algorithm (HEiC) which can be used for that purpose. The context and details for this algorithm will be the main focus of the next chapter. There we also show how the results in this chapter can be used to improve the results in [Araya 2019].

2.8 Examples

In this section we detail particular cases where the results of this section are useful. The context we are most interested on is the case of dot product kernels on the sphere, which has connections with geometric graphs. We chose, in addition, to present applications to two widely used kernels: the Gaussian and the polynomial kernels.

2.8.1 Dot product kernels on the sphere

Likewise Section 2.7 we assume that the kernel on the sphere can be represented by a one dimensional function $f(t)$. Given Eq.(2.15), we have the following expression for the eigenvalues of the kernel, which follows from the Rodrigues formula [Szego 1939, eq. 4.3.1]

$$\lambda_l^* = a_{l,d} b_{l,d} \int_{-1}^{1} f(t) \frac{d^l}{dt^l} \varrho_{\gamma+l}(t) dt \tag{2.17}$$

where $b_{l,d} = \frac{(-1)^l}{2^l l!} \frac{(2d-2)^{(l)}}{\left(\frac{d-1}{2}\right)^{(l)}}$ and $a_{l,d} = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \frac{l!}{(2d-2)^{(l)}}$.

2.8.1.1 Polynomial kernel

We consider a function f of the form $f(t) = \frac{1}{2}(1+t)^q$, which a popular choice as a kernel function (see for example [Rasmussen 2006, Ch.4]). Using integration by parts iteratively in (2.17) we obtain

$$\lambda_l^* = \begin{cases} a_{l,d} b_{l,d} (q)_l \int_{-1}^{1} (1+t)^{q-l} \varrho_{\gamma+l}(t) dt & \text{for } l \leq q \\ 0 \quad \text{otherwise} \end{cases}$$

Expanding the binomial term we obtain, for $l \leq q$

$$\lambda_l^* = a_{l,d} b_{l,d} (q)_l \sum_{j=0}^{q-l} \binom{q-l}{j} \int_{-1}^{1} t^j \varrho_{\gamma+l}(t) dt \tag{2.18}$$

For $j < q-l$ we have

$$\begin{aligned} \int_{-1}^{1} t^j \varrho_{\gamma+l}(t) dt &= \int_{-1}^{1} t^j (1-t^2)^{\gamma+l-1/2} dt \\ &= \frac{1}{2} \operatorname{Beta}(\frac{j}{2} + 1/2, \gamma + l + 1/2) \end{aligned}$$

Using this we obtain for $l \leq q$

$$\begin{aligned} \lambda_l^* &= \frac{1}{2} a_{l,d} b_{l,d} (q)_l \sum_{j=0}^{q-l} \binom{q-l}{j} \operatorname{Beta}(\frac{j}{2} + \frac{1}{2}, \gamma + l + 1/2) \\ &= \frac{(-1)^l (q)_l \Gamma(\frac{d}{2})}{2^{l+1}\sqrt{\pi}} \sum_{j=0}^{q-l} \binom{q-l}{j} \frac{\Gamma(\frac{j}{2} + \frac{1}{2})}{\Gamma(\frac{j}{2} + \frac{d-2}{2} + l)} \end{aligned}$$

Observe that given that $\lambda_l^* = 0$ for $l \geq q$ the kernel is finite rank and the hypothesis of Theorems 3 and 4 are satisfied.

We can also use Cor.15, but observe that, even if $f(\cdot)$ is finite rank, we still need an estimation of the decrease rate of λ_l^*, which is specially important for higher values of q. Given the relation between $\{\lambda_l^*\}_{l\geq 0}$ and $\{\lambda_i\}_{i\geq 1}$ we can deduce the decay rate. Using the properties of the gamma function, it can be proved that λ_l^* is decreasing, but to find a tight bound for each λ_l^* is more involved. On the other hand, to have an idea of value of the scaling factor for the bound in Theorem 3, we can compute the eigenvalues numerically. In Figure 2.1 we show an example of the decrease of the coefficients λ_l^* polynomial kernel on the sphere $\mathbb{S}^7$ for different values of q. Note that the larger eigenvalue has polynomial growth, which is a consequence of the fact that $\max_{-1\leq t\leq 1} f(t) = 2^{q-1}$. Given the large difference in the values between the larger and the smaller eigenvalues, the scaling factor will a play a big role in the estimation of $|\lambda_i(T_n) - \lambda_i|$ when using any of the main results of this chapter.

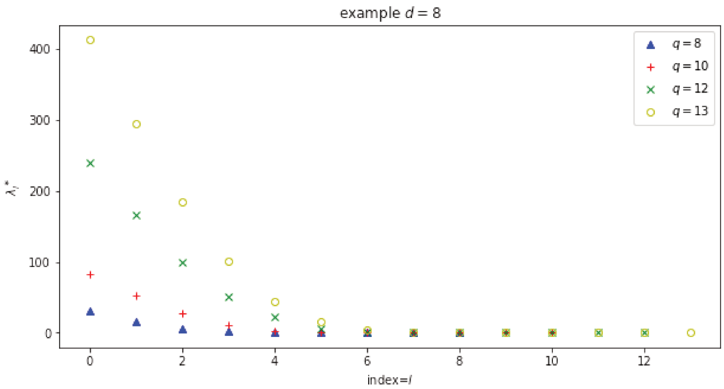

Figure 2.1: Values of λ_l^* for the polynomial kernel in $\mathbb{S}^7$.

2.8.1.2 Constant and linear graphons

Since smooth graphons on the sphere can be conveniently approximated by series of Gegenbauer (ultraspherical) polynomials, we describe here what their spectrum looks like in the finite rank case.

Likewise to Section 2.7, we consider $\gamma = \frac{d-2}{2}$. We start with the *constant graphon* with is the related to the first polynomial in the Gegenbauer basis which is $G_0^\gamma(t) = 1$. More specifically, we consider $W_1(x,y) = p_0 G_0^\gamma(\langle x,y\rangle) = p_0$, where $p_0 \in [0,1]$, which is a rank 1 graphon. This coincides with the well-known *Erdös-Rényi* graphon. If we generate a graph with this model, following Section 2.7.1, with $\{X_i\}_{1\leq i\leq n}$ a uniform sample on the sphere, then the probability that X_i and X_j are connected for any $i,j \in \{1,\cdots,n\}$ is p_0. That is, for any two nodes the probability that they are connected is the same, regardless of their position on the sphere. For this reason, this model can be considered as structureless (see [Bubeck 2016]). Its eigenvalues are (we use the decreasing indexing)

$$\begin{cases} {\lambda_1}^{(1)} & = p_0 \\ {\lambda_i}^{(1)} & = 0, \text{for all } i > 1 \end{cases}$$

In this case, the eigenvalue ${\lambda_1}^{(1)}$ which has multiplicity one has a clear interpretation in the context of graphon theory. Indeed, if we consider the normalized degree we have

$$d_K(x) = \int_{\mathbb{S}^{d-1}} p_0 d\sigma(y) = p_0$$

thus the non-zero eigenvalue $\lambda_1^{(1)}$ is just the mean degree of the graphon (which asymptotically will be the mean degree of the generated graph). We note $T_n^{(1)}$ the kernel matrix associated with W_1. Observe that in this case we can apply Theorem 3, for instance, and given the multiplicity of λ_1 we obtain

$$|\lambda_1(T_n^{(1)}) - {\lambda_1}^{(1)}| \lesssim \frac{p_0}{\sqrt{n}} \sqrt{\log d/\alpha}$$

with probability larger than $1 - \alpha$. For all $i > 1$ we have $\lambda_i(T_n^{(1)}) = 0$. We now consider the graphon $W_2(x, y) = p_0 G_0^\gamma(\langle x, y\rangle) + p_1 G_1^\gamma(\langle x, y\rangle) = p_0 + p_1 2\gamma\langle x, y\rangle$, which has rank $1 + d_1$, where the d_1 is the first spherical harmonic space dimension given in (2.11). It is easy to see that $d_1 = d$. This graphon is based on the first two Gegenbauer (ultraspherical) polynomials $G_0^\gamma(t) = 1$ and $G_1^\gamma(t) = 2\gamma t$ and we call it *linear graphon*. The eigenvalues for this model are given by

$$\begin{cases} {\lambda_1}^{(2)} & = p_0 \\ {\lambda_2}^{(2)} & = p_1 \frac{d-2}{d} \\ {\lambda_i}^{(2)} & = 0, \text{for all } i > 2 \end{cases}$$

The eigenvalue ${\lambda_1}^{(2)}$ has multiplicity one and the eigenvalue ${\lambda_2}^{(2)}$ has multiplicity d. At first glance ${\lambda_2}^{(2)}$ is $O(1)$, but since by definition a graphon takes values $0 \leq W_2(x, y) \leq 1$, the values p_0 and p_1 must satisfy certain constraints. In this particular case, we see that $p_0 \in [0, 1]$ and $p_0 \pm p_1 2\gamma \geq 0$ so $|p_1| \leq \frac{p_0}{2\gamma}$. That implies that ${\lambda_2}^{(2)}$ is decreasing on d. More specifically, since $\gamma = \frac{d-2}{2}$ we have ${\lambda_2}^{(2)} = O(\frac{1}{d})$. As we saw in Section 2.7, here we have $\mathcal{V}(d + 1) = d + 1$. Applying Theorem 3 we obtain, for $\alpha \in (0, 1)$ and with probability bigger than $1 - \alpha$.

$$|\lambda_1(T_n^{(2)}) - {\lambda_1}^{(2)}| \lesssim \frac{p_0}{\sqrt{n}} \sqrt{(d + 1) \log d/\alpha}$$

$$|\lambda_2(T_n^{(2)}) - {\lambda_2}^{(2)}| \lesssim \sqrt{\frac{\log d/\alpha}{nd}}$$

For all $i > 2$ we have $\lambda_i(T_n^{(2)}) = 0$.

Here we see clearly how the relative concentration inequality improves the accuracy with respect to a simple application of absolute Weyl-type inequalities. More specifically, for the eigenvalue ${\lambda_1}^{(2)}$ we get a better dimensional dependence. If we apply Weyl inequality, we will obtain for ${\lambda_1}^{(2)}$ the same order of concentration that for ${\lambda_0}^{(2)}$, that is $(\frac{1}{\sqrt{n}}\sqrt{2(d+1)\log d/\alpha})$. Using Theorem 3 instead we get $(\frac{1}{\sqrt{nd}}\sqrt{\log d/\alpha})$, which is much better.

As a side note, we see that as the dimension d increases, the eigenvalue ${\lambda_2}^{(2)}$ tends to 0. Using Weyl inequality gives a bound that deteriorates in the dimension, such as the one for $\lambda_1(T_n^{(2)})$ above. On the other hand, the relative concentration bound perform better having a scaling term that decrease with the dimension (and in the other term the dimension increase but only logarithmically), giving a much better picture of what actually happens.

2.8.1.3 Proximity and logistic graphons

We consider the graphon $W_g(x,y) = \mathbf{1}_{\langle x,y\rangle \geq 0}$, which in the dense setting is equivalent to angular version of the classic random geometric graph with threshold parameter $\tau = 0$(see [Penrose 2003] or [Bubeck 2017, Sect.2]). We will use the names *proximity graphon* or *threshold graphon* indistinctly. If we generate a random graph by the model described in Section 2.7.1, with $\{X_i\}_{i\in\mathbb{N}}$ a uniform sample on the sphere, then the corresponding nodes X_i and X_j will be connected if and only if they belong to the same semi-sphere. Applying (2.15) we get

$$\lambda_l^* = a_{l,d}\int_0^1 G_l^\gamma(t)\varrho_\gamma(t)dt = a_{l,d}b_{l,d}\int_0^1 \frac{d^l}{dt^l}\varrho_{\gamma+l}(t)dt$$

where $a_{l,d} = \frac{c_l b_d}{d_l}$ and $b_{l,d} = \frac{(-1)^l}{2^l l!}\frac{(2d-2)^{(l)}}{\left(\frac{d-1}{2}\right)^{(l)}}$. The computations for this case are similar, but more involved than in Sect. 2.8.1.1. Similar to that case, it is easier to describe the eigenvalues following the spherical harmonics order, using the $*$ notation

$$\begin{cases} \lambda_0^* & = \frac{1}{2} \\ \lambda_i^* & = 0, \text{ for } i > 0 \text{ even} \\ \lambda_i^* & = \frac{(-1)^{l+\lceil l/2\rceil}}{2\pi}\,\text{Beta}\left(\frac{d}{2},\frac{l}{2}\right), \text{ for } i \text{ odd} \end{cases}$$

where $\text{Beta}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the classic Beta function. Since this function is neither regular nor finite rank we cannot apply directly Theorems 3 and 4. Indeed, from the expression for eigenvalues λ_l^* we deduce that for d fixed, asymptotically as l tends to infinity

$$|\lambda_l^*| \sim \Gamma\big(\frac{d}{2}\big)\cdot\big(\frac{l}{2}\big)^{-\frac{d}{2}}$$

using the Stirling asymptotic approximation of the Beta function.

Clearly the eigenvalues do not fulfill hypothesis H. Indeed, by (2.11) the series with term $|\lambda_l^*|d_l$ is not summable. Nevertheless, we can apply the results to the m-fold composition of the operator $T_{W_g}^{\circ m}$ with $m \in \mathbb{N}$, which is an integral operator with kernel:

$$W_g^{\circ m}(x,y) = \int_{(\mathbb{S}^{d-1})^{m-1}} W_g(x,z_1)\cdot W_g(z_1,z_2)\cdots W_g(z_{m-1},y)d\sigma(z_1)\cdots d\sigma(z_{m-1})$$

where $(\mathbb{S}^{d-1})^{m-1}$ is the $m-1$ product space of the d-dimensional unit sphere. In the context of graphons, the m-fold composition carries important combinatorial information about the family of graphs its represent. For instance, the 2 fold composition gives the number of paths between two nodes and the 3 fold composition is related with the number of triangles [Lovasz 2012][Chap.7].

It is well known that $\lambda(T_{W_g}^{\circ m}) = \{\lambda_i^m\}_{i\geq 1}$ and that the eigenfunctions are the same as T_W. In other words, the following L^2 expansion holds

$$W_g^{\circ m}(x,y) = \sum_{k=0}^{\infty} \lambda_k^m \phi_k(x)\phi_k(y)$$

Taking $m \geq 2$ and using the previous estimation, we have that

$$|\lambda_l^*|^m \sim \left(\frac{l}{2}\right)^{-\frac{dm}{2}}$$

This implies that $|\lambda_i|^m = (i^{\frac{dm}{d-2}})$ by the correspondence between the decreasing and the $*$ indexations. Thus, $W_g^{\circ m}$ satisfies the hypothesis H and H$_1$. In the case $m = 2$, the kernel matrix is

$$(T_n^{\circ 2})_{ij} := \frac{1}{n}\int_{\mathbb{S}^{d-1}} W_g(X_i,z)W_g(z,X_j)d\sigma(z)$$

and we have $|\lambda_i|^2 = O(l^{-\frac{d}{d-2}})$ and $\mathcal{V}_1(i) = O(i)$ by Lemma 13. The previous implies that $W_g^{\circ 2} \in \mathcal{W}_2^p([-1,1],\varrho')$ with $p = 1$. Using Cor. 15 we obtain, with probability higher than $1-\alpha$.

$$|\lambda_i^2 - \lambda_i(T_n^{\circ 2})^2| \lesssim_\alpha i^{-\delta+\frac{\beta\delta}{2(\delta-1)}} n^{-1/2}$$

with $\delta = -\frac{1}{2} - \frac{1}{d-2}$.

We consider the logistic graphon $W_{lg}(x,y) = f(\langle x,y\rangle)$, where $f(t) := \frac{e^{rt}}{1+e^{rt}} = \frac{1}{1+e^{-rt}}$. This model was introduced in [Hoff 2002] and since then many variants have appeared. The symmetry with respect to $\frac{1}{2}$ of the logistic function, implies by (2.15) that the eigenvalues of W_{lg} are given by

$$\lambda_l^* = a_{l,d}\int_{-1}^{1} f(t)G_l^\gamma(t)\varrho_\gamma(t)dt = a_{l,d}b_{l,d}\int_0^1 \frac{1-e^{-rt}}{1+e^{-rt}}\frac{d^l}{dt^l}w_{\gamma+l}(t)dt$$

The eigenvalues of $T_{W_{lg}}$ depend on r in such a way that when $r = 0$ the spectrum of $T_{W_{lg}}$ coincide with the spectrum of the constant graphon with parameter $p_0 = 1/2$ and when $r \to \infty$ the spectrum of $T_{W_{lg}}$ converge to the spectrum of T_{W_g}. We can regard the logistic graphon model as an interpolation between the constant (Erdös-Rényi) graphon and the proximity (geometric) graphon. It is interesting to note that when $r = 0$ the rank of W_{lg} is one and when $r > 0$, W_{lg} has infinite rank. It is easy to see that for $r > 0$ we will have roughly the same problem that in the previous case, as the eigenvalues will not satisfy the asymptotic decay conditions in the definition H. This is, again, a manifestation of the fact the operator associated to W_{lg} is Hilbert-Schmidt, but not trace-class. Using the square operator $T_{W_{lg}}^{\circ 2}$ we obtain a similar result that in the previous case for $r > 0$ for the eigenvalues, which results in a slower rate.

2.8.2 Gaussian kernel

First we consider a kernel used in the context of Gaussian regression [Zhu 1998]. In the one dimensional version, we take $\Omega = \mathbb{R}$ with a measure μ with density with respect to the Lebesgue measure $d\mu(x) = \frac{1}{\sqrt{\pi}}e^{-x^2}dx$ and $K(x,y) = e^{-\frac{1}{2}x^2-\frac{1}{4}(x-y)^2-\frac{1}{2}y^2}$. Its eigenvalues and normalized eigenfunctions are given in [Zhu 1998][sec.4] (see also [Fasshauer 2011, sec. 6.2]), which in the unidimensional case are, for $k \in \mathbb{N}$

$$\lambda_k = \frac{2^{-2k}}{(\frac{1}{2}(1+\sqrt{2})+\frac{1}{4})^{k+\frac{1}{2}}} \leq \frac{2}{5^{k+\frac{1}{2}}}$$

$$\phi_k(x) = \frac{\sqrt[8]{2}}{\sqrt{2^k k!}} \exp\big(-\frac{x^2}{\sqrt{2}}\big) H_k(\sqrt[4]{2}x)$$

where $H_k(\cdot)$ is the k-th order Hermite polynomial (see [Szego 1939, Ch.5]). We note that the eigenvalues have an exponential decreasing rate. On the other hand, using the results in [Indritz 2019] we have for all x

$$\exp\big(-\frac{x^2}{\sqrt{2}}\big) H_k(\sqrt[4]{2}x) \leq \sqrt{2^k k!}$$

Thus $\|\phi_k\|_\infty \leq \sqrt[8]{2}$. Consequently, the hypothesis H_2 for Theorem 4 holds with $s = 0$ and $\delta = \log 5$. We apply Theorem 4, obtaining with probability larger than $1-\alpha$

$$|\lambda_i(T_K) - \lambda_i(T_n)| \lesssim_\alpha e^{-i\log 5} n^{-1/2} \leq e^{-1.6i} n^{-1/2}$$

where T_n is the normalized kernel matrix.

Now, we consider the kernel $K_2(x,y) = e^{-\frac{1}{4}(x-y)^2}$ with the same Ω and μ. It is well known (see[Zhu 1998][sec.4]) that the eigenvalues are the same as the case of K above. The L^2 normalized eigenfunctions are [Fasshauer 2011, sec. 6.2]

$$\phi_k(x) = \frac{\sqrt[8]{2}}{\sqrt{2^k k!}} \exp\big(-(\sqrt{2}-1)\frac{x^2}{2}\big) H_k(\sqrt[4]{2}x)$$

Notice that also in this case the functions also have a uniform bound (which is larger than in the previous case) and the same result applies. That is

$$|\lambda_i(T_{K_2}) - \lambda_i(T_n)| \lesssim_\alpha e^{-1.6i} n^{-1/2}$$

with probability larger than $1-\alpha$.

2.9 Mathematical tools

Here we sum up the tools from relative perturbation and the concentration inequalities used in the proofs of Theorems 3 and 4 and the propositions and lemmas in Section 2.4.

2.9.1 Perturbation results

The following eigenvalue perturbation theorem is due to Ostrowski [Horn 2012, Thm.4.5.9] and [Braun 2005, Cor.3.54]

Theorem 16. *Let $A \in \mathbb{R}^{n\times n}$ be a Hermitian matrix and $S \in \mathbb{R}^{n\times n}$ be a nonsingular matrix. Then for each $1 \leq i \leq n$ there exists $\theta_i > 0$ such that*

$$\lambda_i(SAS^*) = \theta_i \lambda_i(A)$$

In addition, it holds

$$|\lambda_i(SAS^*) - \lambda_i(A)| \leq |\lambda_i(A)| \|S^*S - \mathrm{Id_n}\|_{op}$$

Remark 5. *The previous theorem is also valid for S singular [Horn 2012, Cor.4.5.11].*

The previous theorem can be extended to the case where S is not necessarily a square matrix [Braun 2005, Cor.3.59]

Corollary 17. *Let $A \in \mathbb{R}^{n\times n}$ be a Hermitian matrix and $S \in \mathbb{R}^{d\times n}$ matrix then*

$$|\lambda_i(SAS^*) - \lambda_i(A)| \leq |\lambda_i(A)| \|S^*S - \mathrm{Id_n}\|_{op}$$

From the previous result we deduce the following corollary

Corollary 18. *Under the same conditions of Corollary 17 we have*

$$\|SAS^* - A\|_F \leq \|A\|_F \|S^*S - \mathrm{Id_n}\|_{op}$$

2.9.2 Concentration inequalities

In this chapter we use some classic one dimensional concentration inequalities, such as Hoeffding and Bernstein inequalities (see [Boucheron 2013, sec.2]) to control the tail of the random variable majorizing $\|E_R\|_{op}$.

We also use the Bernstein's theorem [Tropp 2012, Thm.6.1] throughout this chapter.

Theorem 19 (Matrix Bernstein)**.** *Let $\{S_j\}_{1\leq j\leq n}$ a sequence of independent, self-adjoint, random matrices of dimension d. We assume that $\mathbb{E}(S_j) = 0$ and $\|S_j\|_{op} \leq L$ a.s. Defining $\sigma^2 = \|\sum_{1\leq j\leq n} \mathbb{E}(S_j^2)\|_{op}$, we have*

$$\mathbb{P}(\|\sum_{1\leq k\leq n} S_j\|_{op} \geq t) \leq d \exp \frac{-t^2/2}{\sigma^2 + Lt/3}$$

Furthermore, we have

$$\mathbb{E}(\|\sum_{1\leq j\leq n} S_j\|_{op}) \leq \sqrt{2\sigma^2 \log d} + \frac{L}{3} \log d$$

Another important tool used is the concentration inequalities for U-statistics that we used to control quantities related to the residual matrix E_R. The specific result we used was first proven in [Gine 2000], using the Massart's version of Talagrand's concentraion inequality, and later in [Houdré 2003] using a slightly different method that provides explicit constants. Here we present the result for canonical kernels, which are kernels where the expectation with respect to one variable is equal to the expectation with respect to the other (see [Arcones 1993] or [Gine 2015] for a formal definition and on how to decompose a kernel in sum of canonical kernels). The following theorem is formulated in the decoupled version (with two sets of random variables, instead of one), but passing to the regular (undecoupled) kernels is standard, using the results in [De La Pena 1995].

Theorem 20. *Let $h_{i,j}$ be a bounded canonical kernel and $\{X_i^{(1)}\}_{1\leq i\leq n}, \{X_j^{(2)}\}_{1\leq j\leq n}$ two set of independent random variables. Then, there exists a constant $L > 0$ such that*

$$\mathbb{P}(|\sum_{i,j\leq n} h_{i,j}(X_i^{(1)}, X_i^{(2)})| \geq x) \leq L \exp\big[-\frac{1}{L}\min\{\frac{x^2}{C^2}, \frac{x}{D}, \frac{x^{2/3}}{B^{2/3}}, \frac{x^{1/2}}{A^{1/2}}\}\big]$$

where

$$A = \max_{ij} \|h_{ij}\|_\infty, \quad C = \mathbb{E}\sum_{ij} h_{ij}^2$$

$$B^2 = \max\Big\{\|\sum_i \mathbb{E}h_{ij}^2(X_i^{(1)}, y)\|_\infty, \|\sum_i \mathbb{E}h_{ij}^2(x, X_i^{(2)})\Big\}$$

$$D = \sup\Big\{\mathbb{E}\sum_{ij} h_{ij}(X_i^{(1)}, X_j^{(2)}) f_i(X_i^{(1)}) g_j(X_j^{(2)}),\ s.t\ \mathbb{E}\sum_i f_i^2(X_i^{(1)}) \leq 1, \mathbb{E}\sum_i g_i^2(X_i^{(2)}) \leq 1\Big\}$$

2.10 Extensions and improvements

Here we discuss alternatives to the use of the Matrix Bernstein theorem, which can result in a better bound. The fact that the matrix Φ_R has independent rows, opens the possibility to use some tighter results such as [Vershynin 2012a, Prop. 2.1], [Vershynin 2012b, Thm. 5.39] or [Koltchinskii 2017, Cor. 2].

It is easy to see that the rows of Φ_R are sub-Gaussian. Indeed, from the fact that λ_i, ϕ_i are an eigenvalue, eigenfunction couple, we have

$$\|\phi_i\|_\infty \leq \frac{\sup_{x,y\in\Omega}|K(x,y)|}{\lambda_i}$$

The fact that the functions $\{\phi_i\}_{1\leq i\leq R}$ are bounded implies that each row $\xi_j := (\phi_1(X_j), \cdots, \phi_R(X_j))$ is sub-Gaussian, because the L^∞-norm dominates the sub-

Gaussian norm $\|\cdot\|_{\psi_2}$. Moreover, we have

$$\begin{aligned}\|\xi_i\|_{\psi_2} &= \sup_{\|u\|\leq 1} \|\langle \xi_i, u\rangle\|_{\psi_2}\\ &\lesssim \sup_{\|u\|\leq 1} \|\langle \xi_i, u\rangle\|_{\infty}\\ &\lesssim \|\sum_{k=1}^{R'} \phi_k^2\|_{\infty}^{1/2} = \sqrt{\mathcal{V}_1(R)} \qquad (2.19)\end{aligned}$$

By [Koltchinskii 2017, Cor. 2] we have in this case: with probability larger than $1-e^{-t}$:

$$\|\frac{1}{n}\Phi_R\Phi_R^T - \mathrm{Id}_R\|_{op} \lesssim \sqrt{\frac{R}{n}} \vee \frac{R}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n} \qquad (2.20)$$

The previous inequality is essentially sharp, as signaled in [Koltchinskii 2017]. A more general result is discussed in [Koltchinskii 2017], where the dimension R is replaced by the effective rank $r(A) = \frac{\mathrm{Tr}(A)}{\|A\|_{op}}$, where A is a matrix (or more generally a self adjoint trace-class operator). Those results can be extended to the more general context of Hilbert spaces (see [Koltchinskii 2017]). Note however that in those results, the $\|\cdot\|_{\psi_2}$-norm of the random vectors in consideration are treated as constant in those results and the sub-Gaussian norm is hidden in the notation $\lesssim$. Those results are better adapted for random vectors where the $\|\cdot\|_{\psi_2}$-norm do not grow with the dimension. This is not the case for the columns in Φ_R, in general, because we can not assert that the eigenfunctions are uniformly bounded, even for very regular kernels [7]. Both theorems are proven using the same principles (mainly a chaining argument), but the one in [Koltchinskii 2017] uses a more powerful generic chaining inequality which results in a better dependency in $\|\Phi_R\|_{\psi_2}$.

On the other hand, the result of Prop. 6 can be rewritten as

$$\|\Phi_R\Phi_R^T - \mathrm{Id}_R\|_{op} \lesssim \mathcal{V}_1(R)\Big[\frac{\log R + t}{n} \vee \sqrt{\frac{\log R + t}{n}}\Big] \qquad (2.21)$$

with probability larger than $1-e^{-t}$. As we already mentioned, we can not compare (2.21) with (2.20) as readily written, because of the constants hidden in (2.20). On the other hand, results in [Koltchinskii 2017] are derived from more general tail bounds obtaining by generic chaining technique, which are expressed directly in terms of the $\|\cdot\|_{\psi_2}$-norm.

Before turning our attention to those results, we need a few definitions. We will say that a centered square integrable[8] random vector X in $\mathbb{R}^R$ is pre-Gaussian if there exists a centered Gaussian random vector Y in $\mathbb{R}^R$, such that $Cov(Y) = Cov(X)$. We need to define the so-called γ_2 functional to state the generic chaining tails bound Theorem 22 below. The presentation of this material is borrowed from

[7] There is, in fact, an example credited to Smale, published by Zhou in [Zhou 2002], of a positive definite $\mathcal{C}^\infty$ kernel whose eigenfunctions exhibit an explosive behavior in L^∞-norm.

[8] in its most general form this definition is formulated for weekly integral random variables in a Banach space. See [] for details.

[Koltchinskii 2017] and the more classic treatment can be found in [Ledoux 1991]. We consider $N_0 = 1$ and $N_n = 2^{2^n}$, for $n \geq 1$. For a given metric space $(\mathcal{T}, d)$ we consider Δ_n a sequence of increasing partitions. We say that Δ_n is admissible if $\mathrm{card}(\Delta_n) \leq N_n$. Let $\Delta_n(t)$ be the only set in Δ_n that contains t and define

$$\gamma_2(\mathcal{T}, d) = \inf_{\Delta_n \text{ admissible}} \sup_{t\in\mathcal{T}} \sum_{n=0}^{\infty} 2^{n/2} D(\Delta_n(t))$$

The following is the celebrated *majorizing measures theorem* due to M. Talagrand.

Theorem 21 ([Talagrand 1996]). *Let X_t, $t \in \mathcal{T}$ be a centered Gaussian process. Define*

$$d(s,t) = \sqrt{\mathbb{E}\big((X_s - X_t)^2\big)}, \quad \textit{for } s,t \in \mathcal{T}$$

Then $\exists K > 0$ and absolute constant such that

$$\gamma_2(\mathcal{T}, d) \leq K\mathbb{E}\sup_{t\in\mathcal{T}} X_t$$

The following tail bound was obtained by Dirksen [Dirksen 2015, Cor.5.7] and independently by Bednorz [Bednorz 2014, Thm.1].

Theorem 22. *Let $X_1, \cdots, X_n$ be i.i.d random variables on (Ω, μ) and $\mathcal{T}$ a class of measurable functions defined on Ω. For all $t > 1$ we have with probability larger than $1 - e^{-t}$*

$$\sup_{f\in\mathcal{T}} |\frac{1}{n}\sum_{i=1}^{n} f^2(X_i) - \mathbb{E}f(X_i)^2| \lesssim \tilde{\mathcal{V}}(f)\frac{\gamma_2(\mathcal{T},\psi_2)}{\sqrt{n}} \vee \frac{\gamma_2^2(\mathcal{T},\psi_2)}{n} \vee \tilde{\mathcal{V}}^2(f)\Big(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\Big) \quad (2.22)$$

where $\tilde{\mathcal{V}}(f) = \sup_{f\in\mathcal{T}} \|f\|_{\psi_2}$.

We will use Theorem 22 on the class of linear functions defined by

$$\mathcal{T} = \{f_u(x) = \langle x, u\rangle : \ u \in \mathbb{S}^{R-1}, \} \quad (2.23)$$

that is the projections onto a given direction $u \in \mathbb{R}^R$. In the case of the random vectors ξ_i we have

$$\sup_{f\in\mathcal{T}} \|f\|_{\psi_2} \leq \sqrt{\mathcal{V}_1(R)}$$

Observe that Theorem 21 offers a bound on $\gamma_2(\mathcal{T}, d)$, where d depends on the $L^2(\mathbb{R}^R)$ norm. In this part, we deal with Gaussian processes rather than our original distribution. As we have already discussed, for a Gaussian random vector, we have the equivalence of L^2 and ψ_2 norms, in the sense that if X is a Gaussian random vector in $\mathbb{R}^R$ then

$$c\|\langle X, u\rangle\|_{L^2} \leq \|\langle X, u\rangle\|_{\psi_2} \leq C\|\langle X, u\rangle\|_{L^2}$$

for two absolute constants $c, C > 0$. This is turns imply that

$$\gamma_2(\mathcal{T}, \psi_2) \lesssim \gamma_2(\mathcal{T}, d)$$

Now we use Theorem 4, to obtain

$$\begin{aligned}\gamma_2(\mathcal{T}, d) &\leq \mathbb{E} \sup_{f \in \mathcal{T}} X_f \\ &= \mathbb{E} \sup_{u \in \mathbb{S}^{R-1}} X_u \\ &= \mathbb{E} \sup_{u \in \mathbb{S}^{R-1}} |\langle X, u \rangle| \\ &\lesssim \mathbb{E}\|X\| = \sqrt{R}\end{aligned}$$

where in the second to last equality we use the fact that we can replace a Gaussian process indexed in the sphere by the projections of a Gaussian vector on the sphere. It is easy to see that, in the case of the random vectors $\xi_1, \cdots, \xi_n$, eq. (2.22) yields, with probability larger that $1 - e^{-t}$

$$\|\frac{1}{n}\Phi_R\Phi_R^T - \mathrm{Id}_R\|_{op} \lesssim \tilde{\mathcal{V}}(R)\sqrt{\frac{R}{n}} \vee \frac{R}{n} \vee \tilde{\mathcal{V}}(R)^2\Big(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\Big) \tag{2.24}$$

where $\tilde{\mathcal{V}}(R) = \sup_{f \in \mathcal{T}} \|f\|_{\psi_2} = \sup_{\|u\| \leq 1} \|\langle \xi, u \rangle\|_{\psi_2}$. From (2.19) we see that (2.24) will be an improvement over (2.21) in certain cases. Indeed, take for example the case $\mathcal{V}_1(R) \geq R$ and notice that in (2.24) there will not appear the $\log R$ term under the square root.

Removing this logarithmic term seems a little improvement, but the approach we followed in this section, suggest also a different parametrization for the problem, using the sub-Gaussian norm, instead of the $\|\cdot\|_\infty$, since the latter could be much larger than the former. Unfortunately, computing the sub-Gaussian norm $\|\cdot\|_{\psi_2}$, or find a tight bound, might not be as straightforward as to estimate the $\|\cdot\|_\infty$ norm. However, there have been research on that front in the recent years [Arbel 2017, Arbel 2019] and more is expected to come.

2.11 Some proofs

Here we gather proofs for some selected results. More details can be found in the appendix of [Araya 2020].

Proof of Prop. 6. We use the matrix Berstein theorem. We note that

$$\Phi_R^T\Phi_R - \mathrm{Id}_R = \sum_{j=1}^{n}(Z_jZ_j^T - \frac{1}{n}\mathrm{Id}_R)$$

where $Z_j \in \mathbb{R}^R$ is given by $(Z_j)_k = \frac{1}{\sqrt{n}}\phi_k(X_j)$ By definition $\mathbb{E}[Z_jZ_j^T] = \frac{1}{n}\mathrm{Id}_R$. It is easy to prove that

$$\|Z_jZ_j^T - \frac{1}{n}\mathrm{Id}_R\|_{op} \leq \frac{|\mathcal{V}_1(R) - 1|}{n}$$

and

$$\|\sum_{k=1}^{n}\mathbb{E}[(Z_jZ_j^T - \frac{1}{n}\mathrm{Id}_R)^2]\|_{op} \leq \frac{|\mathcal{V}_1(R) - 1|}{n}$$

Using Theorem 19 with $S_j = Z_j Z_j^T - \frac{1}{n}\mathrm{Id}_R$, $d = R$, $L = \frac{|\mathcal{V}_1(R)-1|}{n}$ and $\sigma^2 = L$ we get

$$\mathbb{P}(\|\Phi_R^T\Phi_R - \mathrm{Id}_R\|_{op} \geq t) \leq R \exp \frac{-nt^2}{2\mathcal{V}_1(R)(1+\frac{t}{3})}$$

From this the $1-\alpha$ confidence version is direct. □

Proof of Lemma 8. From Prop.6 we have that

$$\mathbb{P}\Big(\|\Phi_R^T\Phi_R - \mathrm{Id}_R\|_{op} \lesssim \sqrt{\frac{\mathcal{V}_1(R)\log R/\alpha}{n}}\Big) \geq 1-\alpha$$

We put $\tau = \sqrt{\frac{\mathcal{V}_1(R)\log R/\alpha}{n}}$ and solving for τ gives the result. □

Proof of Prop. 10. The first inequality comes from (**??**). Indeed, when $\tau_{n,R,\alpha} < 1$ using Prop. 7 we have

$$\|(\Phi_R^T\Phi_R)^{-1}\Phi_R^T E_R \Phi_R (\Phi_R^T\Phi_R)^{-1}\|_{op} \lesssim_\alpha \gamma_1(n,R)$$

with probability larger than $1-\alpha$. Using the previous and Prop. 6 in (**??**) we obtain

$$|\lambda_i(T_n) - \lambda_i(M)| \lesssim_\alpha \big(|\lambda_i| + \gamma_1(n,R)\big)\sqrt{\frac{\mathcal{V}_1(R)\log R}{n}}$$

By the assumption $\lambda_i > \gamma_2(n,R)$ and the block diagonal structure of M we obtain that $\lambda_i(M) = \lambda_i$.

In the case $|\lambda_i| < \gamma_2(n,R)$, we cannot assure that $\lambda_i(M) = \lambda_i$, but we know that $\lambda_i(M)$ is at distance at most $\gamma_2 - |\lambda_i| < \gamma_2$ from λ_i. On the other hand, we use Prop. 7, the submultiplicative property of $\|\cdot\|_{op}$ and the fact that projection has norm 1 to obtain

$$\|P_1 E_R P_1\|_{op} \lesssim_\alpha \gamma_2(n,R)$$

then using (2.6) we obtain the desired result. □

Proof of Prop. 11. From $T_n = \Phi_R\Lambda_R\Phi_R^T + E_R$ we see that $|\lambda_i(T_n)| < \gamma_2(n,R)$, because $\lambda_i(\Phi_R\Lambda_R\Phi_R^T) = 0$. On the other hand, by definition of $\gamma_2(n,R)$(because it contains the tail b_R) we have $|\lambda_i| \leq \gamma_2(n,R)$. Then $|\lambda_i(T_n) - \lambda_i| \leq \gamma_2(n,R)$ with the required probability. □

Proof skerch of Prop. 7. The idea is to use the Thm. 20 to prove (2.7). Indeed, $\Phi_R^T E_R \Phi_R$ is a $R \times R$ matrix and the entry (i,j) can be written as $\frac{1}{n^2}\sum_{l_1,l_2=1}^{n} \sum_{k>R} \lambda_k \phi_i(X_{l_1})\phi_k(X_{l_1})\phi_k(X_{l_2})\phi_j(X_{l_2})$.

For each entry we apply Thm. 20. We need to check that the kernel is canonic (which follows from the orthogonality of the functions ϕ_i and ϕ_k for $i \leq R$ and $k \geq R$) and find bound for each of the parameters A, B, C and D defined in Thm. 20.

Once we have a bound for each term of $\Phi_R^T E_R \Phi_R$, we use the matrix norm bound $\|\cdot\|_{op} \leq \sqrt{R}\|\cdot\|_1$ and the result follows.

To prove (2.8) we use that the left hand side is controlled by $\|E_R\|_{op}$. Given that $(E_R)_{ij} = \sum_{k>R} \lambda_k \phi_k(X_i)\phi_k(X_j)$, we can split E_R in two terms as $E_R = E_R^+ - E_R^-$, considering the positive in E_R^+ and negative eigenvalues in E_R^-. By the triangle inequality is sufficient to control both terms, that is $\|E_R\|_{op} \leq \|E_R^+\|_{op} + \|E_R^-\|_{op}$. Given that E_R^+ and E_R^- are p.s.d matrices, both operator norms are bounded by their traces. For $\mathrm{Tr}(E_R^+)$ we can use the law of large numbers, given that the sum of the diagonal of E_R^+ is a sum of independent random variables, to prove that $\mathrm{Tr}(E_R^+) \to b_R^+ = \sum_{i>R,\ \lambda_i>0} |\lambda_i|$. Then using the Hoeffding inequality, we get the result for the positive part. The negative part is analogous and the result follows by adding both the positive and the negative parts.

□

Proof sketch of Thm. 3. We divide in two cases: $\lambda_i > b_{i+m_i}$ and $\lambda_i \leq b_{i+m_i}$. For the first, we use Prop. 10, that with confidence at least $1-\alpha$

$$|\lambda_i(T_n) - \lambda_i(T_W)| \lesssim \gamma_1(n, i+m_i) + |\lambda_i + \gamma_1(n, i+m_i)|\sqrt{\frac{\mathcal{V}_1(i+m_i)\log{(i+m_i)/\alpha}}{n}}$$

We choose n_0 to be minimum integer such that $\gamma_1(n, i+m_i) < \lambda_i$ (this condition generates a non empty set, given that $\gamma_1(n, i+m_i) \to 0$) and the desired conclusion follows. The case $\lambda_i \leq b_{i+m_i}$ is more involved, since includes a refinement of the bound in Prop. 10 by iterating the same argument. The choose of the value n_0 depends on the spectral gap quantity $\mathrm{Gap}_+(j) = \min_{l>j:\lambda_j\neq\lambda_l} |\lambda_j - \lambda_l|$. □

Proof sketch of Thm. 4. The idea is for each index i to use either Prop. 10 or Prop. 11 (the one delivering the tighter bound). We can see the results in this section as finding a rule that tell us how to select the truncation parameter R best adapted for each i. For H_1, for instance, we prove that

$$\gamma_1(n, R) = \mathcal{O}(R^{2s+\frac{5}{2}-\delta}n^{-1}) = \mathcal{O}(n^{\beta\delta'(2s+\frac{5}{2}-\delta)-1})$$
$$\gamma_2(n, R) = \mathcal{O}(n^{\beta\delta'(1-\delta)}) + \mathcal{O}(n^{\beta\delta'(2s+1-\delta)-\frac{1}{2}})$$

Then we check that for $\beta = \mathcal{O}(\frac{1}{s})$ choosing $\delta' = \frac{\delta}{\delta-1}$ we can use Prop. 10 and get

$$|\lambda_i(T_n) - \lambda_i| \lesssim_\alpha i^{-\delta+(s+1/2)\beta\frac{\delta}{\delta-1}}n^{-1/2}$$

with probability larger than $1-\alpha$. For the rest of β's we use Prop.11 and the orders of γ_1 and γ_2 to prove

$$|\lambda_i(T_n) - \lambda_i| \lesssim_\alpha i^{-\delta+(2s+1)\frac{\delta}{\delta-1}}n^{-1/2}$$

with probability larger than $1-\alpha$. The proves for H_2 and H_3 are in the same lines. □

Proof of Lemma 13. We prove it for the case where i can be decomposed exactly as $i = \sum_{l \in I} d_l$. In this case, it is direct from (2.13) that the sum $\sum_{j=1}^{i} \|\phi_j^2\|_\infty$ is equal to $\sum_{l \in I} |Z_l(x,x)| = i$. Given that $\mathcal{V}_1(i)$ increasing, this is enough to prove its linear order. □

Proof of Cor. 5. It is clear that we only need to check that the term that depends on i is summable. For instance, if K satisfy H$_2$ we have (given that $n = i^{1/\beta}$) with probability larger than $1 - \alpha$

$$\begin{aligned}
\sum_{i=1}^{n} |\lambda_i(T_n) - \lambda_i| &\lesssim_\alpha \sum_{i}^{n} e^{-\delta i + g_2(s+1/2)\log i} n^{-1/2} \\
&\lesssim_\alpha e^{-\delta i + (2s+1)\log i} n^{-1/2} \\
&\lesssim_\alpha n^{-1/2}
\end{aligned}$$

where we used that $i \geq \log i$ and that the series with exponentially decreasing term is summable. On the other hand, we have the following convexity inequality

$$\delta_2(\lambda(T_n), \lambda(T_K)) \leq \sum_{i=1}^{n} |\lambda_i(T_n) - \lambda_i| + b_n$$

Given H$_2$ we have $b_n = \mathcal{O}(e^{(1-\delta)n})$, then it holds $b_n = \mathcal{O}(n^{-1/2})$ for n larger than $\frac{n}{\log n} \geq \frac{1}{2(\delta-1)}$. Inserting this into the previous inequality we get

$$\delta_2(\lambda(T_n), \lambda(T_K)) \lesssim_\alpha n^{-1/2}$$

The cases H$_3$ and H$_1$(with the additional assumption) are analogous.

□

Chapter 3

Latent distance estimation for RGG's on the sphere.

Contents

3.1 Introduction

In this chapter we study the problem of estimating latent distances in the random geometric graph (RGG) model. The material presented here is based mainly on the article [Araya 2019], but we also expand its content by reporting more numerical experiments and theoretical improvements we discovered after its publication.

The RGG model is a latent space model, based on the existence of latent points which we assume are randomly placed in a metric space. The connection between these points is utterly determined by their position. Probably, the most recurrent example of this graph family is the random *τ-proximity graph* where two nodes are connected if their associated latent points are at distance τ or smaller for a $\tau > 0$.

This define a determinstic connection rule and the only source of randomness lies in the random placement of the latent points (which we typically assume are i.i.d with a common law μ). We often call this model the classical RGG model, because was probably the first random graph model in this family to be introduced, in [Gilbert 1961]. A more general model was introduced in [Hoff 2002] for the study of social networks, where the latent points are sampled similarly as before, but the connection rule is now non-deterministic: the probability that an arc between two nodes exists depends on the distance between the corresponding latent points. Since then many variants and generalizations have been introduced.

The fact that there is an underlying geometry to the RGG model has made it popular in many application such as wireless networks modeling [Jia 2004], sensor localization [Li 2009, Eren 2017], protein interactions [Higham 2008a], link prediction [Sarkar 2010], Physics [Cunningham 2017] and social networks [Hoff 2002]. This underlying geometry confers to these graphs the property of *homophily* which is explained as " the principle that a contact between similar people occurs at a higher rate than among dissimilar people"[McPherson 2001]. This feature is present in many real world networks, whereas it is not present in some simpler yet ubiquitous models such as the Erdös-Rényi model. For that reason, the problems dealing with recovering structures from the observation of this type of networks has become increasingly relevant in statistics and machine learning. A few examples on this direction are the problem of detecting the presence of an underlying geometry in [Bubeck 2016] and community detection on Euclidean random graphs [Abbe 2017].

We focus on the problem of estimating latent distances from a single observation of a simple graph (without loops and with no weights on the arcs) for angular RGG. In that model, the underlying space is the Euclidean sphere and points are sampled according to the uniform (surface) measure. We place ourselves in the dense case, where the probability of connection between two nodes will be determined by the inner product between them, that is it will be represented by a dot product kernel (as in Section 2.7, from the previous chapter). The nice representation of the kernel spectral expansion coming from harmonic analysis will allow us to construct an estimator of the Gram matrix (from which the distances can be readily be derived) based on a set of eigenvectors of the adjacency matrix.

There are two main type of algorithms for the latent distance estimation. On one hand, spectral methods is one family with contains the works in [Tang 2013] and ours, for example. Another different line of research considers the estimation given by the graph theoretic distance (the distance between two nodes is the length of its shortest path), which is often regularized using a Semi-Definite Programming(SDP) approach. Examples on this line are [Diaz 2018] and [Arias-Castro 2018].

It is worth mentioning that there are related problems and variants of latent distance estimation. We briefly mention some of them. One related problem is the sensor network localization [Oh 2010](which sometimes uses a latent distance estimation as first step). Noisy versions of the latent distance recovery and the latent point localization also exists [Javanmard 2013], where we assume that we observe a randomly perturbed version of the true graphs (typically due to measurement errors).

3.2 Random geometric graph via graphon model

We describe the generative model for dense networks used in this chapter, which is a generalization of the classical random geometric graph model introduced by Gilbert in [Gilbert 1961]. We base our definition on the W-random graph model described in [Lovasz 2012, Sec. 10.1], which uses the graphon formalism. The central objects will be graphon functions on the sphere, which are kernel functions of the form $W : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [0,1]$. Throughout this Chapter, we consider an underlying measure space $(\mathbb{S}^{d-1}, \sigma)$, where σ is the uniform measure on the sphere. On $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ we consider the product measure $\sigma \times \sigma$.

To generate a simple graph from a graphon function, we first sample n points $\{X_i\}_{i=1}^n$ independently on the sphere $\mathbb{S}^{d-1}$, according to the uniform measure σ. These are the so-called *latent points*. Secondly, we construct the matrix of distances between these points, called the *Gram matrix* $\mathcal{G}^*$ (we will often call it population Gram matrix) defined by

$$\mathcal{G}^*_{ij} := \langle X_i, X_j \rangle$$

and the so-called *probability matrix*

$$\Theta_{ij} = W_n(X_i, X_j) = \rho_n W(X_i, X_j)$$

which is also a $n \times n$ matrix. Given that for $x, y \in \mathbb{S}^{d-1}$ we have that the spherical distance between x, y is $\arccos \langle x, y \rangle$ and the Euclidean distance is $2 - 2\langle x, y \rangle$, then estimating the latent distances reduces to estimating the Gram matrix.

The function W gives the precise meaning for the “link” function, because it determines the connection probability between X_i and X_j. One difference with the previous chapter is the introduction of the scale parameter $0 < \rho_n \leq 1$, which allow us to control the edge density of the sampled graph given a function W, see [Klopp 2017a] for instance. The case $\rho_n = 1$ corresponds to the dense case (the parameter Θ_{ij} do not depend on n) and when $\rho_n \to 0$ the graph will be sparser. Our main results will hold in the regime $\rho_n = \Omega(\frac{\log n}{n})$, which we call *relatively sparse*. As in the previous chapter, we will work with the normalized version of the probability matrix $T_n := \frac{1}{n}\Theta$. If there exists a function $f : [-1,1] \to [0,1]$ such that $W(x,y) = f(\langle x, y \rangle)$ for all $x, y \in \mathbb{S}^{d-1}$ we will say that W is a geometric graphon.

Finally, we define the random *adjacency matrix* $\hat{T}_n$, which is a $n \times n$ symmetric random matrix that has independent entries (except for the symmetry constraint $\hat{T}_n = \hat{T}_n^T$), conditional on the probability matrix, with laws

$$n(\hat{T}_n)_{ij} \sim \mathcal{B}(\Theta_{ij})$$

where $\mathcal{B}(m)$ is the Bernoulli distribution with mean parameter m. Since the probability matrix contains the mean parameters for the Bernoulli distributions that define the random adjacency matrix it has been also called the *parameter matrix* [Chatterjee 2015]. Observe that the classical RGG model (or proximity graph) on the sphere is a particular case of the described W-random graph model when $W(x,y) = \mathbb{1}_{\langle x,y\rangle \geq \tau}$. In that case, since the entries of the probability matrix only have values in $\{0,1\}$, the adjacency matrix and the probability matrix are coincident. Depending on the context, we use $\hat{T}_n$ for the random matrix as described above or for an instance of this random matrix, that is for the adjacency matrix of the observed graph. This will be clear from the context.

It is worth noting that graphons can be, without loss of generality, defined in $[0,1]^2$. The previous affirmation means that for any graphon there exists a graphon in $[0,1]^2$ that generates the same distribution on graphs for any given number of nodes. However, in many cases the $[0,1]^2$ representation can be less revealing than other representations using a different underlying space. This is illustrated in the case of the *prefix attachment* model in [Lovasz 2012, example 11.41], where the graphon in $\big([0,1]\times[0,1]\big)^2$ is a simple function, but the equivalent graphon in $[0,1]^2$ is a complicated fractal function.

In the sequel we use the notation $\lambda_0, \lambda_1, \cdots, \lambda_{n-1}$ for the eigenvalues of the normalized probability matrix T_n. Similarly, we denote by $\hat{\lambda}_0, \hat{\lambda}_1, \cdots, \hat{\lambda}_{n-1}$ the eigenvalues of the matrix $\hat{T}_n$. We recall that T_n (resp. $\hat{T}_n$) and $\frac{1}{\rho_n}T_n$ (resp.$\frac{1}{\rho_n}\hat{T}_n$) have the same set of eigenvectors. We will denote by v_j for $1 \leq j \leq n$ the eigenvector of T_n associated to λ_j, which is also the eigenvector of $\frac{1}{\rho_n}T_n$ associated to $\frac{1}{\rho_n}\lambda_j$. Similarly, we denote by $\hat{v}_j$ to the eigenvector associated to the eigenvalue $\rho_n\hat{\lambda}_j$ of $\hat{T}_n$. Given that a geometric graphon is essentially a bounded dot product kernel defined on $\mathbb{S}^{d-1}$, we adopt the notation $\{\lambda_l^*\}_{l\geq 0}$ for the eigenvalues of the integral operator T_W as in Section 2.7. We recall that in the $*$ notation, the indexation of the eigenvalues follows the degree of the spherical harmonics and not the non-increasing order of their absolute values.

3.3 Geometric graphon eigensystem

Here we cover asymptotic and finite sample results for the eigenvalues and eigenfunctions of the three central objects involved: the integral operator T_W, the normalized probability (kernel) matrix T_n and the normalized adjacency matrix $\hat{T}_n$. There is some intersection with the material covered in Section 2.7, but with a slight reformulation given the scale parameter ρ_n. The Sobolev regularity parameter δ we use below correspond to the one discussed in the previous chapter, Section 2.6.

The following are the key spectral asymptotic and finite sample results which are relevant in our approach to Gram matrix reconstruction:

- The spectrum of $\frac{1}{\rho_n}T_n$ converges a.s. to the spectrum of T_W in the $\delta_2(\cdot,\cdot)$ metric. We saw in Chapter 2 that this is a consequence of [Koltchinskii 2000, Thm.1].

- The spectral projections of T_n converge to the spectral projections of T_W. In [Koltchinskii 1998] We have

$$\sup_{f,g\in\mathcal{F}} |\langle P_i^{\varepsilon}(T_n)\tilde{f},\tilde{g}\rangle_{L^2(\sigma_n)} - \langle P_i(T_W)f,g\rangle_{L^2(\sigma)}| \to 0$$

where P_i^ε represents the orthogonal projection of the eigenspace generated by a set of eigenvalues in ε-*cluster* around λ_i, $\tilde{f}$ is the vector $\big(f(X_1), \cdots, f(X_n)\big)$ and σ_n is the empirical measure associated to $\{X_i\}_{1\le i\le n}$. The class of functions $\mathcal{F}$ must satisfy certain technical conditions (of the Glivenko-Cantelli type), but since we do not use the result directly we do not enter into details and refer the interested reader to [Koltchinskii 1998]. In our main results, spectral gap conditions related to the notion of ε-cluster (which will be explain with more detail in the next section) will be used and a finite sample version of this result will be stated.

- For W with Sobolev regularity δ, we have with probability larger than $1-\alpha$:

$$\delta\big(\lambda(\frac{1}{\rho_n}T_n, \lambda(T_W))\big) \lesssim_\alpha \left(\frac{\log n}{n}\right)^{\frac{\delta}{2\delta+d-1}} \tag{3.1}$$

  This was proved in [De Castro 2020].

- Matrices $\hat{T}_n$ approach to matrix T_n in operator norm as n gets larger, which is a consequence of the results in [Bandeira 2016], which describe the concentration properties of the spectral norm for a matrix with independent entries. More specifically, we apply [Bandeira 2016][Cor.3.3] to the centered matrix $Y = \hat{T}_n - T_n$ we get

$$\mathbb{E}(\|\hat{T}_n - T_n\|_{op}) \lesssim \frac{\sqrt{D_0}}{n} + \frac{\sqrt{D_0^*}\sqrt{\log n}}{n} \tag{3.2}$$

  where $D_0 = \max_{0\le i\le n}\sum_{j=1}^n \Theta_{ij}(1-\Theta_{ij})$ and $D_0^* = \max_{ij}\|Y_{ij}\|_\infty$. We clearly have that $D_0 = \mathcal{O}(n\rho_n)$ and $D_0^* \le 1$, which implies that

$$\mathbb{E}\|\hat{T}_n - T_n\|_{op} \lesssim \max\left\{\frac{\rho_n}{\sqrt{n}}, \frac{\sqrt{\log n}}{n}\right\}$$

  We see that this inequality do not improve if ρ_n is smaller than in the relatively sparse case, that is $\rho_n = \Omega(\frac{\log n}{n})$. We prove in Theorem 32 that, as a corollary of the results in [Bandeira 2016], we have

$$\frac{1}{\rho_n}\|\hat{T}_n - T_n\|_{op} \le_{\alpha/4} C\max\{\frac{1}{\sqrt{\rho_n n}}, \frac{\sqrt{\log n}}{\rho_n n}\} \tag{3.3}$$

  An analogous bound can be obtained for the Frobenius norm replacing $\hat{T}_n$ with $\hat{T}_n^{\text{usvt}}$ the USVT estimator defined in [Chatterjee 2015]. For our main results, Proposition 25 and Theorem 26 the operator norm bound will suffice.

One of main results is concentration inequality relating the eigenvectors of $\hat{T}_n$ with the sampled eigenfunctions of T_W, which is a key step in our method, in light of the material presented below.

Our algorithm for estimating the Gram matrix will make use of a reconstruction formula, that follows from the harmonic representation of the graphon function. We recall from Section 2.7 that in the spherical context the eigenfunctions of T_W are the spherical harmonics in $\mathbb{S}^{d-1}$. As we saw, the dimension of each eigenspace of T_W is fixed and corresponds to d_k, the dimension of the k-th spherical harmonic space, which we recall from (2.11) satisfies $d_0 = 1, d_1 = d$ and $d_k = \binom{k+d-1}{k} - \binom{k+d-3}{k-2}$.

From the addition theorem (see eq. (2.13)) we have that

$$\sum_{j=d_{k-1}}^{d_k} \phi_j(x)\phi_j(y) = c_k G_k^\gamma(\langle x, y \rangle)$$

where G_k^γ are the Gegenbauer polynomials of degree k with parameter $\gamma = \frac{d-2}{2}$ and $c_k = \frac{2k+d-2}{d-2}$. The Gegenbauer polynomial of degree one (linear) is $G_1^\gamma(t) = 2\gamma t$ (see [Dai 2013, Appendix B2]), hence we have $G_1^\gamma(\langle X_i, X_j \rangle) = 2\gamma \langle X_i, X_j \rangle$ for every $1 \leq i, j \leq n$. In consequence, by the addition theorem we have the following reproducing formula

$$G_1^\gamma(\langle X_i, X_j \rangle) = \frac{1}{c_1} \sum_{k=1}^{d} \phi_k(X_i)\phi_k(X_j) \tag{3.4}$$

where we recall that $d_1 = d$. This implies the following relation for the population Gram matrix, observing that $2\gamma c_1 = d$

$$\mathcal{G}^* := \frac{1}{n}(\langle X_i, X_j \rangle)_{i,j} = \frac{1}{2\gamma c_1} \sum_{j=1}^{d} v_j^* v_j^{*T} = \frac{1}{d} V^* V^{*T} \tag{3.5}$$

where v_j^* is the $\mathbb{R}^n$ vector with i-th coordinate $\phi_j(X_i)/\sqrt{n}$ and V^* is the matrix with columns v_j^*. In a similar way, we define for any matrix U in $\mathbb{R}^{n \times d}$ with columns $u_1, u_2, \cdots, u_d$, the matrix $\mathcal{G}_U := \frac{1}{d} U U^T$. As part of our main theorem we prove that for n large enough there exists a matrix $\hat{V}$ in $\mathbb{R}^{n \times d}$ where each column is one of the eigenvector of $\hat{T}_n$, such that $\hat{\mathcal{G}} := \mathcal{G}_{\hat{V}}$ approximates $\mathcal{G}^*$ well, in the sense that the norm $\|\hat{\mathcal{G}} - \mathcal{G}^*\|_F$ converges to 0 at a rate which is that of the nonparametric estimation of a function on $\mathbb{S}^{d-1}$.

3.4 Eigenvalue gap assumption

In this section we describe one of our main hypotheses on W, needed to ensure that the space $\text{span}\{v_1^*, v_2^*, \cdots, v_d^*\}$ can be effectively recovered with the vectors $\hat{v}_1, \hat{v}_2, \cdots, \hat{v}_d$ using the algorithm presented in Section 3.6. Informally, we assume that the eigenvalue λ_1^*, associated to $G_1^\gamma(t)$, is sufficiently isolated from the rest of the spectrum of T_W (not counting multiplicity). Given a geometric graphon W, we define the *spectral gap* of W relative to the eigenvalue λ_1^* by

$$\text{Gap}_1(W) := \min_{j \neq 1} |\lambda_1^* - \lambda_j^*|$$

which quantifies the distance between the eigenvalue λ_1^* and the rest of the spectrum. In particular, we have the following elementary proposition.

Proposition 23. *It holds that* $\mathrm{Gap}_1(W) = 0$ *if and only if either exists* $j \neq 1$ *such that* $\lambda_j^* = \lambda_1^*$, *or* $\lambda_1^* = 0$.

Proof. Observe that the unique accumulation point of the spectrum of T_W is zero. The proposition follows from this observation. □

To recover the population Gram matrix $\mathcal{G}^*$ with our Gram matrix estimator $\hat{\mathcal{G}}$ we require the spectral gap $\Delta^* := \mathrm{Gap}_1(W)$ to be different from 0, which will ensure the identiafibility of the set of d eigenvectors that reconstruct the Gram matrix in (3.5). This assumption has been made before in the literature[1], mainly because some version of the Davis-Kahan $\sin\theta$ theorem (see for instance [Chatterjee 2015], [Levin 2017], [Tang 2013]) is used. More precisely, our results will hold on the following event

$$\mathcal{E} := \Big\{ \delta_2\Big(\lambda\big(\frac{1}{\rho_n}T_n\big), \lambda(T_W)\Big) \vee \frac{2^{\frac{9}{2}}\sqrt{d}}{\rho_n \Delta^*}\|T_n - \hat{T}_n\|_{op} \le \frac{\Delta^*}{4} \Big\},$$

for which we prove the following

Lemma 24. *Assume that* $\Delta^* > 0$, *then there exists* $n_0 \in \mathbb{N}$ *such that for* $n \geq n_0$ *and for* $\alpha \in (0,1)$ *we have with probability larger than* $1-\alpha$

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\alpha}{2}$$

Remark 6. *The value* n_0 *will depend on* W *and* α. *This dependence can be made explicit using* (3.10) *and* (3.7)

$$\max\Big\{\sqrt{\frac{\rho_n}{n}}, \frac{\sqrt{\log n}}{n}\Big\} \leq \frac{\Delta^{*2}}{2^{15/2}C\sqrt{d}} \quad \text{and} \quad \frac{\log n}{n} \leq \Big(\frac{\Delta^*}{8C'}\Big)^{\frac{2\delta+d-1}{\delta}}$$

where $C, C' > 0$. *This will be clearer from the proof sketch in Section 3.10.*

3.5 Latent distance estimation

The following theorems are the main results of the article [Araya 2019]. We sketch their proofs in Section 3.10.

Proposition 25. *On the event* $\mathcal{E}$, *there exists one and only one set* Λ_1, *consisting of* d *eigenvalues of* $\hat{T}_n$, *whose diameter is smaller than* $\rho_n\Delta^*/2$ *and whose distance to the rest of the spectrum of* $\hat{T}_n$ *is at least* $\rho_n\Delta^*/2$. *Furthermore, on the event* $\mathcal{E}$, *our algorithm (Algorithm 1) returns the matrix* $\hat{\mathcal{G}} = (1/c_1)\hat{V}\hat{V}^T$, *where* $\hat{V}$ *has by columns the eigenvectors corresponding to the eigenvalues on* Λ_1.

Theorem 26 ([Araya 2019])**.** *Let* W *be a regular geometric graphon on* $\mathbb{S}^{d-1}$ *with regularity parameter* δ *and such that* $\Delta^* > 0$. *Then there exists a set of eigenvectors* $\hat{v}_1, \cdots, \hat{v}_d$ *of* $\hat{T}_n$ *such that*

$$\|\mathcal{G}^* - \hat{\mathcal{G}}\|_F = O(\Delta^{*-1} n^{-\frac{\delta}{2\delta+d-1}})$$

[1] Mainly in the context of matrix estimation and manifold learning.

where $\hat{\mathcal{G}} = \mathcal{G}_{\hat{V}}$ and $\hat{V}$ is the matrix with columns $\hat{v}_1, \cdots, \hat{v}_d$. Moreover, this rate is the minimax rate of nonparametric estimation of a regression function f with Sobolev regularity s in dimension $d-1$.

The condition $\Delta^* > 0$ allow us to use Davis-Kahan type results for matrix perturbation to prove Theorem 26. With this and concentration for the spectrum we are able to control with high probability the terms $\|\hat{\mathcal{G}} - \mathcal{G}\|_F$ and $\|\mathcal{G} - \mathcal{G}^*\|_F$. Theorem 26 proves that our proposed estimator is consistent under the spectral gap condition and, in addition, that it achieves the same rate that non-parametric estimation (we refer the interested reader to [Emery 1998, Chp.2] and [De Castro 2020]). The precise rate here is a consequence of the use of Davis-Kahan theorem, where the deviation of the eigenvectors is largely determined by the deviation of the eigenvalues.

As we mentioned in the introduction to this chapter, the spectral gap condition cannot be removed in general, this can be seen from the case of the Erdös-Rényi random graph, where the graphon has only zero eigenvalues except for λ_0^*, see Section 2.8.1.2. In that case it is easy to see that any sample configuration $\{X_i\}_{1\leq i\leq n}$ has the same probability to generate any graph with n nodes. In other words, the position of the points $\{X_i\}_{1\leq i\leq n}$ does not play a role in the determination of the arcs. This is also explicit in of our numerical experiments, where the application of our method to the Erdös-Rényi graph gives poor results.

It is worth noting that in the works [Diaz 2018] and [Arias-Castro 2018], which propose an estimator based of the matrix of graph theoretic distances, related conditions appear. In the first case, the authors consider only the proximity graphon, which has an spectral gap $\text{Gap}_1(\cdot)$ bounded away from 0 (the precise values are reported in Section 2.8.1.3). In the case of [Arias-Castro 2018] the condition is on the increase rate of the graphon function, which has similar consequences with respect to the spectral gap condition.

3.6 Algorithms

We present two algorithms that work under the eigengap assumption: one for the latent distances estimation and the other for estimating the dimension of the latent sphere $\mathbb{S}^{d-1}$.

3.6.1 Estimation of the distances

The Harmonic EigenCluster algorithm(HEiC) (see Algorithm 1 below) receives the observed adjacency matrix $\hat{T}_n$ and the sphere dimension d as its inputs to reconstruct the eigenspace associated to the eigenvalue λ_1^*. In order to do so, the algorithm selects d vectors in the set $\hat{v}_1, \hat{v}_2, \cdots \hat{v}_n$, whose linear span is close to the span of the vectors $v_1^*, v_2^*, \cdots, v_d^*$ defined in Section 3.3. The main idea is to find a subset of $\{\hat{\lambda}_0, \hat{\lambda}_2, \cdots, \hat{\lambda}_{n-1}\}$, which we call Λ_1, consisting on d_1 elements (recall that $d_1 = d$) and where all its elements are close to λ_1^*. This can be done assuming that the event $\mathcal{E}$ defined above holds (which occurs with high probability). Once we have the

Algorithm 1: Harmonic EigenCluster(HEiC) algorithm

```
Input: $(\hat{T}_n, d)$ adjacency matrix and sphere dimension
    $\Lambda^{\text{sort}} = \{\hat{\lambda}_1^{\text{sort}}, \cdots, \hat{\lambda}_{n-1}^{\text{sort}}\}$ ←eigenvalues of $\hat{T}_n$ sorted in decreasing order
    $\Lambda_1 \leftarrow \{\Lambda_1^{\text{sort}}, \cdots, \Lambda_{1+d}^{\text{sort}}\}$: where $\Lambda_i^{\text{sort}}$ is the $i$-th element in $\Lambda^{\text{sort}}$
    Initialize $i = 2$, gap = $\text{Gap}_1(\hat{T}_n; 1, 2, \cdots, d)$
while $i \leq n - d$ do
    if $\text{Gap}_1(\hat{T}_n; i, i+1, \cdots, i+d) >$ gap then
        $\Lambda_1 \leftarrow \{\Lambda_i^{\text{sort}}, \cdots, \Lambda_{i+d}^{\text{sort}}\}$
    end if
    $i \leftarrow i + 1$
end while
Return: $\Lambda_1$, gap
```

set Λ_1, we return the projector onto the span of the eigenvectors associated to the eigenvalues in Λ_1.

For a given set of indices $i_1, \cdots, i_d$ we define

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

and

$$\text{Gap}_1(\hat{T}_n) := \max_{\{i_1, \cdots, i_d\} \in \mathcal{S}_d^n} \text{Gap}_1(\hat{T}_n; i_1, \cdots, i_d)$$

where $\mathcal{S}_d^n$ contains all the subsets of $\{1, \cdots, n-1\}$ of size d. This definition parallels that of $\text{Gap}_1(W)$ for the graphon. Observe any set of indices in $\mathcal{S}_d^n$ will not include 0. Otherwise stated, we claim that we can leave $\hat{\lambda}_0^{\text{sort}}$ out of this definition and it will not be candidate to be in Λ_1. Proposition 36, stated below, will justify this claim. In words, we prove that the largest eigenvalue of the adjacency matrix will be close to the eigenvalue λ_0^* and in consequence can not be close enough to λ_1^* to be in the set Λ_1, given the definition of the event $\mathcal{E}$ and the fact the eigenvalue λ_0^* has multiplicity 1.

To compute $\text{Gap}_1(\hat{T}_n)$ we consider the set of eigenvalues $\hat{\lambda}_j$ ordered in decreasing order. We use the notation $\hat{\lambda}_j^{\text{sort}}$ to emphasize this fact. We define the right and left differences on the sorted set by

$$\begin{aligned}\text{left}(i) &= |\hat{\lambda}_i^{\text{sort}} - \hat{\lambda}_{i-1}^{\text{sort}}| \\ \text{right}(i) &= \text{left}(i+1)\end{aligned}$$

where $\text{left}(\cdot)$ is defined for $1 \leq i \leq n$ and $\text{right}(\cdot)$ is defined for $0 \leq i \leq n-1$. With these definition, we have the following lemma,

Lemma 27. *On the event $\mathcal{E}$, the following equality holds*

$$\text{Gap}_1(\hat{T}_n) = \max\left\{ \max_{1 \leq i \leq n-d-1} \min\{\text{left}(i), \text{right}(i+d)\}, \text{left}(n-d+1) \right\}$$

The set Λ_1 has the form $\Lambda_1 = \{\hat{\lambda}^{\text{sort}}_{i^*}, \hat{\lambda}^{\text{sort}}_{i^*+1}, \cdots, \hat{\lambda}^{\text{sort}}_{i^*+d}\}$ for some $1 \leq i^* \leq n-d-1$. We have that either

$$i^* = \underset{1 \leq i \leq n-d-1}{\arg\max} \; \min\{\text{left}(i), \text{right}(i+d)\}$$

or

$$i^* = n - d$$

depending whether or not one has $\max_{1\leq i \leq n-d-1} \min\{\text{left}(i), \text{right}(i+d)\} > \text{left}(n-d+1)$. The algorithm then constructs the matrix $\hat{V}$ having columns $\{\hat{v}_{i^*}, \hat{v}_{i^*+1}, \cdots, \hat{v}_{i^*+d}\}$ and returns $\hat{V}\hat{V}^T$.

It is worth noting that Algorithm 1's time complexity is $n^3 + n$, where n^3 comes from the fact that it is a spectral algorithm and computing the eigenvalues and eigenvectors of the $n \times n$ matrix $\hat{T}_n$ takes roughly n^3 steps and the linear term is because we explore the whole set of eigenvalues to find the maximum gap for the size d cluster of eigenvalues. In terms of space complexity the algorithm is roughly n^2 because we need to store the matrix $\hat{T}_n$.

Remark 7. *If we change $\hat{T}_n$ in the input of Algorithm 1 to $\hat{T}^{\text{usvt}}_n$ (obtained by the USVT algorithm [Chatterjee 2015]) we predict that the algorithm will give similar results. This is because discarding some eigenvalues bellow a prescribed threshold do not have effect on our method if the threshold is smaller than λ^*_1. However, as preprocessing step the USVT might help in speeding up the eigenspace detection, but this step is already linear in time.*

3.6.2 Estimation of the dimension d

So far we have focused on the estimation of the population Gram matrix $\mathcal{G}^*$. We now give an algorithm, which we called HEiC-dim, to find the dimension d when it is not provided as input. This method receives the matrix $\hat{T}_n$ as input and uses Algorithm 1 as a subroutine to compute a score, which is simply the value of the variable $\text{Gap}_1(\hat{T}_n)$ returned by Algorithm 1. We do this for each d in a set of candidates, which we call $\mathcal{D}$. This set of candidates will be frequently, but not necessarily, fixed to $\{1, 2, 3, \cdots, d_{max}\}$ and the practitioner can choose it differently if additional information about the dimension is available, for instance. Once we have computed a score for each candidate, we pick the candidate that have the maximum score.

Given the guarantees provided by Theorem 26, the previously described procedure will find the correct dimension, with high probability (on the event $\mathcal{E}$) if the true dimension of the ambient sphere is on the candidate set $\mathcal{D}$. This will happen, in particular, when the spectral gap assumptions of Theorem 26 are satisfied.

More formally, we have the following

Corollary 28. *If W in $\mathbb{S}^{d-1}$ is a geometric graphon satisfying $\Delta^* > 0$ and $\mathcal{D}$ is a set of integers containing such that $d \in \mathcal{D}$, then with high probability (on the event $\mathcal{E}$) Algorithm 2 returns the correct dimension.*

```
Algorithm 2: HEiC-dim
    Input: $(\hat{T}_n, \mathcal{D})$
    Run Algorithm 1 for $d \in \mathcal{D}$, $\mathrm{Sc}(d) \leftarrow \mathrm{gap}\big(\mathrm{HEiC}(\hat{T}_n, d)\big)$
    Return: $\arg\max_d(\mathrm{Sc}(d))$
```

Notice that, in contrast with the setting in Algorithm 1, here we do not have a measure of error converging to zero and it is less straightforward to quantify the efficiency. In addition, we have not described how the set $\mathcal{D}$ is explored (in the numerical experiments we do it in increasing order). Indeed, there are many ways in which this algorithm could be more efficient (it is parallelizable for instance), but since we are not pursuing optimal efficiency here, we leave those improvement for future work.

3.7 Numerical experiments and simulations

We generate synthetic data using different geometric graphons. In the first set of examples, we focus in recovering the Gram matrix when the dimension is provided. In the second set we tried to recover the dimension as well.

3.7.1 Recovering the Gram matrix

To measure the error of the algorithm HEiC we will compute each time the mean error(or simply the error in the Frobenius norm), defined by

$$ME_n = \|\hat{\mathcal{G}} - \mathcal{G}^*\|_F$$

which in light of Theorem 26 converges to 0.

We start by considering the graphon $W_1(x, y) = \mathbb{1}_{\langle x,y\rangle \geq 0}$ which defines, through the W-random graph sampling scheme given in Section 3.2, the classical RGG model on $\mathbb{S}^{d-1}$ with threshold 0 (or random proximity graph). Thus any two sampled points $X_i, X_j \in \mathbb{S}^{d-1}$ will be connected if and only if they lie in the same semisphere.

In our first experiment, we fix the dimension to $d = 3$ and consider different values for the sample size n and for each of them we sample 100 Gram matrices and run the Algorithm 1 for each one. In Figure 3.1(left) we show a boxplot for the $\log(ME_n)$ for the different values of n. Notice that the red line in this case is showing how ME_n decrease in terms of n and suggest that a bound of the type $\mathcal{O}(\frac{1}{\sqrt{n}})$ is appropiate for the ME_n. We examine this closer from the theoretical point of view in Section 3.9. In Figure 3.1(right) we see how our algorithm is affected by the change of dimension d. We consider values $d = 3, 5, 7$ and 11 for the dimension and for each n, the ME_n we plot is the mean over the 100 sampled graphs. The results are very similar albeit the spectral gap being smaller as the dimension increase, in which case our bounds deteriorate. In the numerical experiments the bounds are slightly better for larger dimensions. One possible explanation to this fact is that there is

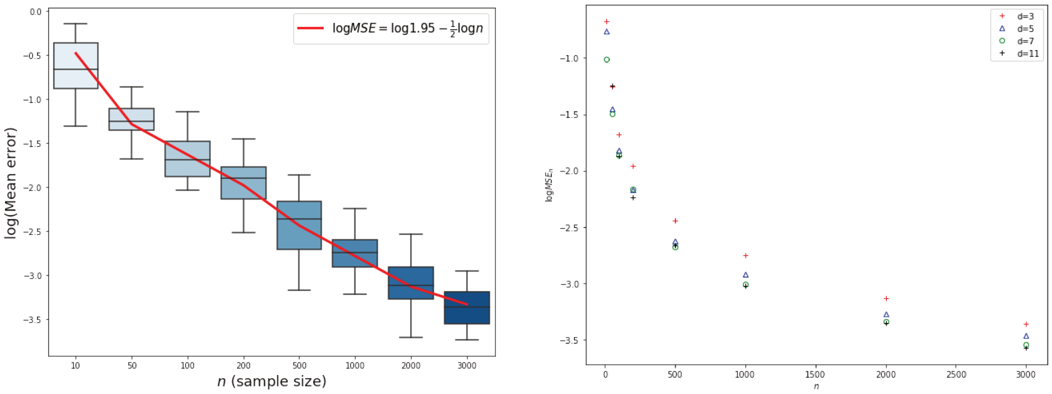

Figure 3.1: In the left we have a boxplot of $\log(ME_n)$ for Algorithm 1 for different values of n for W_1 with $d = 3$. The red line represent a curve fitted for the mean across all repetitions for each n. In the right, we see the $\log(ME_n)$ for W_1 on $\mathbb{S}^{d-1}$ with different values of d.

scaling factor, as in Chapter 2, which compensates the decrease in the spectral gap. The fact that one group of eigenvalues (and eigenvectors) are used in our algorithm would call for the use of the relative concentration results of the previous chapter. However, stated as it is, our proof uses the full spectrum to control the eigenvectors which comes from the form of the Davis-Kahan theorem.

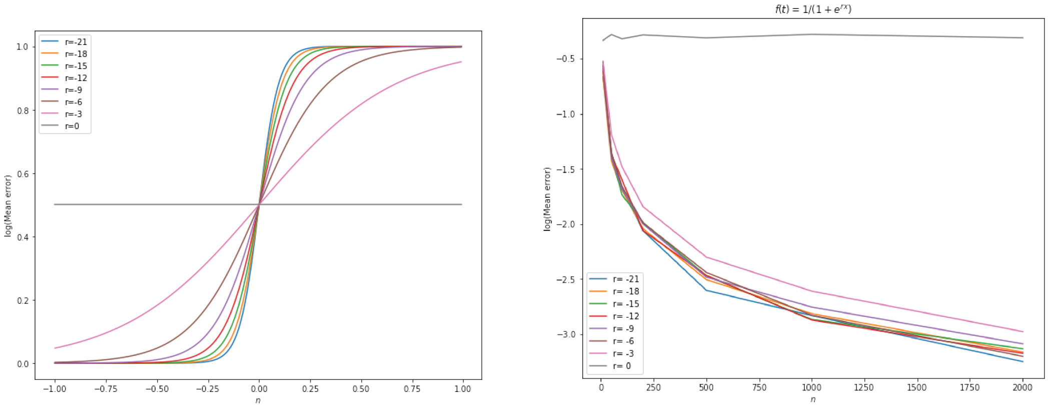

Figure 3.2: In the left we illustrate the logistic graphon for different parameters. In the right, we see the $\log(ME_n)$ for W_2 for different values of r.

Now we consider the logistic graphon $W_2(x,y) = \frac{1}{1+e^{r\langle x,y\rangle}}$, which was discussed in Section 2.8.1.3. This model interpolates between the proximity graphon $W_1(x,y)$ defined above and the Erdös-Rényi graphon $W_0(x,y) = \frac{1}{2}$. In Figure 3.2(left) we plot the graphon functions for different values of r. In Figure 3.2(right) we show the $\log(ME_n)$ for W_2 in each case. Notice that when $r = 0$, we have $W_2 = W_0$. The closer to W_1, in this example, is W_2 with $r = -21$. As we expect the error of the Erdös-Rényi graphon do not decrease with n, since the spectral gap condition is not satisfied and in this case we cannot guarantee reconstruction. As already discussed, in this case no algorithm would achieve reconstruction, since any configuration is equally likely.

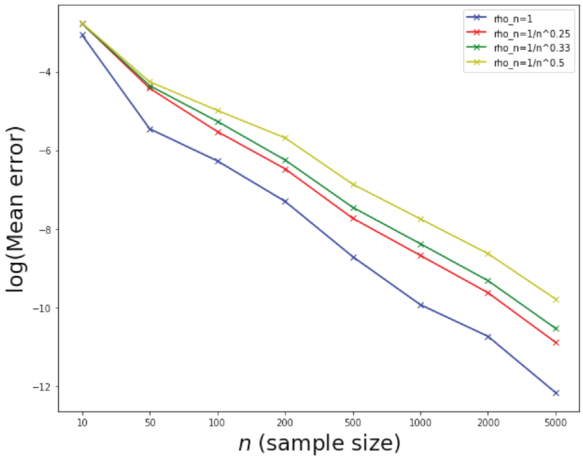

Figure 3.3: $\log(ME_n)$ for W_1 and different values of ρ_n.

We also investigate the effect of the scaling sparsity parameter ρ_n. In Figure 3.3 we plot the $\log(ME_n)$ for the function W_1 for different values of the scale parameter ρ_n. As we might expect, the smaller the ρ_n the slower the error converges to 0, which is due to the fact that the spectral gap is smaller and the dimension fixed. As a general rule, for a fixed d, the smaller the spectral gap the harder the problem becomes, at least asymptotically.

3.7.2 Recovering the dimension d

We conducted a simulation study using graphon W_1, with a sample size $n = 1000$ points on the sphere $\mathbb{S}^{d-1}$ for different values of d and a dimension candidate set of the form $\mathcal{D} = \{1, \cdots, d_{max}\}$. We use Algorithm 2 to recover the underlying dimension d. We repeat this procedure 50 times for each d. In Figure 3.4 we display a boxplot (over the 50 repetitions) for the score of each candidate in $\mathcal{D}$. In the first three cases, that is when $d = 3, 7, 19$ the algorithm can each time differentiate the true dimension from the rest of candidates. In the last case when $d = 77$, the algorithm still peaks in score at the correct dimension, but with a much lower confidence. In our experiments, at $d = 81$ the algorithm cannot distinguish the correct dimension. Notice that the score, which corresponds to the value Gap_1 decrease with the dimension and in the case $d = 77$ is close to 0. In this case, a larger sample size will be needed to recover the dimension. One of the bottlenecks is the computation speed for large sample sizes, given that the time complexity of

HEiC is $\mathcal{O}(n^3)$. Finding ways to speeding up the algorithm will certainly broadens the instances in which could be successfully applied.

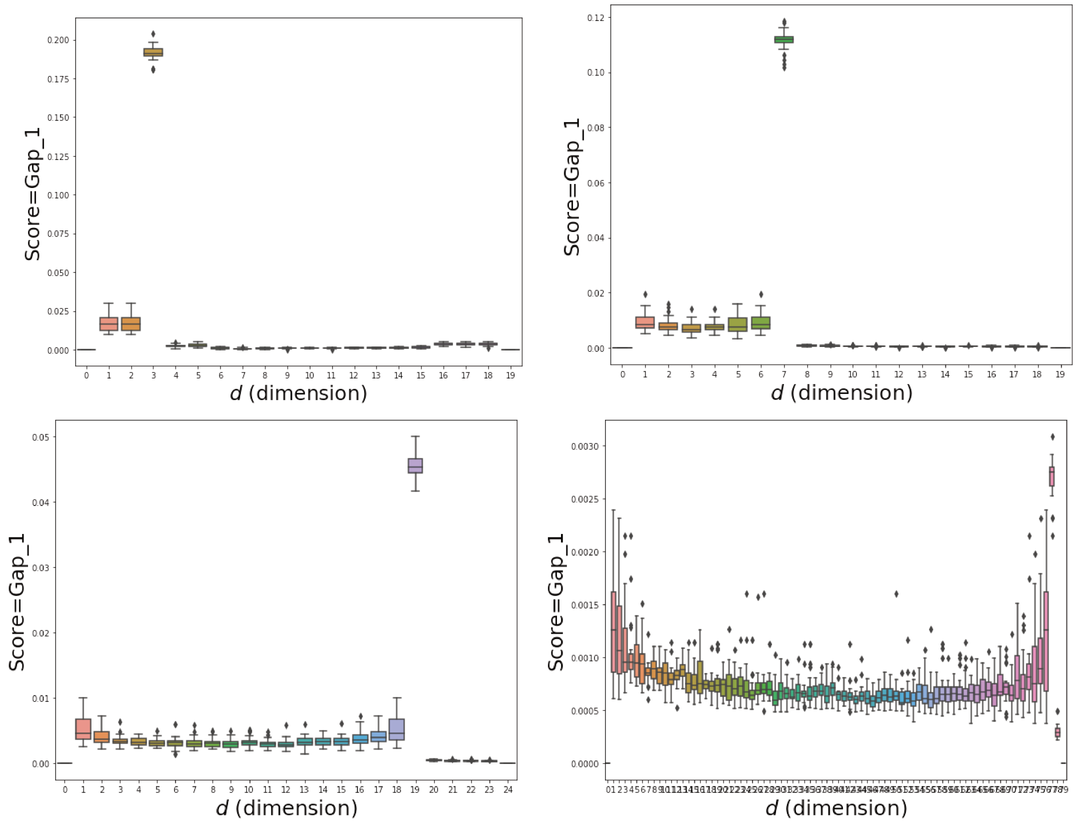

Figure 3.4: Boxplot of the score Gap_1 for Algorithm 2 for different values of d for W_1 with $n = 1000$. In the cases $d = 3, 7, 19$ the algorithm will return the correct dimension.

In Table 3.1 we report running times for Algorithm 1 for reference. Each time correspond to one pass of the algorithm.

3.8 Mathematical tools

3.8.1 Perturbation results

For n large enough, the eigenspace associated to the eigenvalue $\hat{\lambda}_1$ is close to the eigenspace associated to the eigenvalue λ_1. This is precised by the Davis-Kahan *sin* θ theorem. We use the following version which is proved in [Yu 2015]

Theorem 29 (Davis-Kahan)**.** *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* $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \hat{\lambda}_n$ *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* $\lambda_0 := \infty$

n	Time(secs)
20	0.0055
50	0.0278
100	0.044
200	0.128
500	0.818
1000	3.68
2000	16.84
5000	152.52
10000	1379.65

Table 3.1: Running times for the Algorithm 1 in a machine with 3, 3 GHz, Intel i5, 16 Gb RAM.

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 $(v_r, v_{r+1}, \cdots, v_s)$ and $(\hat{v}_r, \hat{v}_{r+1}, \cdots, \hat{v}_s)$ respectively, such that $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$. Then there exists an orthogonal matrix $\hat{O}$ in $\mathbb{R}^{d\times d}$ such that

$$\|\hat{V}\hat{O} - V\|_F \leq \frac{2^{3/2} \min\{\sqrt{d}\|\Sigma - \hat{\Sigma}\|_{op}, \|\Sigma - \hat{\Sigma}\|_F\}}{\min\{\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}\}} \tag{3.6}$$

Also, we need the following perturbation result [Bathia 1997, Thm.VII.2.8]

Theorem 30. *Let A and B two the normal matrices and define $\delta = dist(\lambda(A), \lambda(B))$. If X satisfies the Sylvester equation $AX - XB = Y$, then*

$$\|X\|_F \leq \frac{1}{\delta}\|Y\|_F$$

Another useful perturbation theorem [Bathia 1997, Thm.VII.3.1]

Theorem 31. *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*

$$\|EF\|_F \leq \frac{1}{\delta}\|E(A-B)F\|_F \leq \frac{1}{\delta}\|A - B\|_F$$

3.8.2 Concentration inequalities

The following theorem is a slight reformulation of the [Bandeira 2016, Cor.3.12]

Theorem 32 (Bandeira-Van Handel)**.** *Let Y be a $n \times n$ symmetric random matrix whose entries Y_{ij} are independent centered random variables. There exists a universal constant C_0 such that for $\alpha \in (0,1)$*

$$\mathbb{P}\Big(\|Y\|_{op} \geq 3\sqrt{2D_0} + C_0 D_0^* \sqrt{\log n/\alpha}\Big) \leq \alpha$$

where $D_0 = \max_{0\leq i\leq n} \sum_{j=1}^n \mathbb{E}(Y_{ij}^2)$ and $D_0^* = \max_{ij} \|Y_{ij}\|_\infty$.

Using the previous theorem with $Y = \hat{T}_n - T_n$, which is centered and symmetric, we obtain the tail bound

$$\mathbb{P}\Big(\|\hat{T}_n - T_n\|_{op} \geq \frac{3\sqrt{2D_0}}{n} + C_0 \frac{\sqrt{\log n/\alpha}}{n} \Big) \leq \alpha$$

We use the following result, which can be found in [De Castro 2020]

Theorem 33. *Let W be a graphon on the sphere of the form $W(x, y) = f(\langle x, y\rangle)$. If f belongs to the weighted Sobolev space $Z^s_{w_\gamma}\big((-1, 1)\big)$ then we have*

$$\delta_2(\lambda(\frac{1}{\rho_n} T_n), \lambda(T_W)) \leq_\alpha C \Big(\frac{\log n}{n}\Big)^{\frac{s}{2s+d-1}}$$

where $\leq_\alpha$ means that the inequality holds with probability greater than $1 - \alpha$ for $\alpha \in (0, 1/3)$ and n large enough.

While Theorem 32 gives a bound for the difference of the eigenvalues of the observed matrix with respect to the eigenvalues of the probability matrix, Proposition 33 ensures that the eigenvalues of the empirical matrix are close to these of the integral operator.

Given a set of independent random vectors $X_1, \cdots, X_n$ uniformly distributed on the sphere $\mathbb{S}^{d-1}$ we are interested in the concentration properties of the quantity $\frac{1}{n}\sum_{k=1}^n X_i X_i^T$ around its mean, which is $\mathbb{E}(X_i X_i^T) = \mathrm{Id}_d$ for $1 \leq i \leq n$ (in other words, the vectors X_i are *isotropic*). Since the uniform distribution on the sphere is sub-gaussian [Vershynin 2018, Thm.3.4.6], we can use the following theorem [Vershynin 2012a, Prop.2.1].

Theorem 34. *If $X_1, \cdots, X_n$ are independent random vectors in $\mathbb{R}^d$ with $d \leq n$ which have sub-gaussian distribution. Then for any $\alpha \in (0, 1)$ it holds*

$$\|\frac{1}{n}\sum_{k=1}^n X_k X_k^T - \mathrm{Id}_d\|_{op} \leq_\alpha \sqrt{\frac{d}{n}}$$

3.8.3 Other useful results

To avoid border issues in HEiC algorithm, we use the fact that the eigenvalue λ_0^* associated to the Gegenbauer polynomial $G_0^\gamma(t) = 1$ for $t \in [-1, 1]$ is the largest one. The following results justify this facts.

Lemma 35. *If $W : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [0, 1]$ is such that*

$$W(x, y) = f(\langle x, y\rangle)$$

for $f : [-1, 1] \to [0, 1]$, then

$$d_W(x) := \int_{\mathbb{S}^{d-1}} W(x, y) d\sigma(y)$$

is constant.

Proof. The proof follows from a change of variable. □

The following theorem is an analogous result to a classical theorem of spectral graph theory

Proposition 36. *For a graphon $W : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [0,1]$ we have*

$$\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} W(x,y)d\sigma(x)d\sigma(y) \leq \lambda_0^* \leq \max_{x\in\mathbb{S}^{d-1}} d(x)$$

Since $G_0^\gamma(t) = 1$ we have by Lemma 35 and Prop. 36 that the $\lambda_0^* = \lambda_0^{\text{sort}}$. We make used of the following results from linear algebra

Lemma 37. *Let A, B be two matrices in $\mathbb{R}^{n\times d}$ then*

$$\|AA^T - BB^T\|_F \leq (\|A\|_{op} + \|B\|_{op})\|A-B\|_F$$
$$\|AA^T - BB^T\|_{op} \leq (\|A\|_{op} + \|B\|_{op})\|A-B\|_{op}\,.$$

If it holds that $A^TA = B^TB = I_d$ then

$$\|AA^T - BB^T\|_F \leq 2\|A-B\|_F$$

Lemma 38. *Let B a $n \times d$ matrix with full column rank. Then we have*

$$\|BB^T - B(B^TB)^{-1}B^T\|_F = \|\text{Id}_d - B^TB\|_F$$

3.9 Extensions and improvements

In Chapter 2 we saw that under certain regularity conditions, we can assure the $\mathcal{O}(\sqrt{\frac{1}{n}})$ for the $\delta_2(\cdot,\cdot)$ metric. The previous can be derived from Cor. 5. In the proof we use that for the regularity conditions H_1, H_2 and H_3 we can choose R in such a way that the function $\gamma_2(n,R)$ is $\mathcal{O}(\sqrt{\frac{1}{n}})$. Here the parameter s is equal to 0 (see Section 2.7). In the proof of Theorem 26 we use triangle inequality to bound $\|\mathcal{G}^* - \hat{\mathcal{G}}\|_F$ by the terms $\|\mathcal{G}^* - \mathcal{G}\|_F$ and $\|\mathcal{G} - \hat{\mathcal{G}}\|_F$.

Thanks to Davis-Kahan theorem the term $\|\mathcal{G} - \hat{\mathcal{G}}\|_F$ is already $\mathcal{O}(\frac{1}{\sqrt{n}})$. On the other hand, we used the triangle inequality to express

$$\|\mathcal{G}^* - \mathcal{G}\|_F \leq \|\mathcal{G}^* - \mathcal{G}^*_{proj}\|_F + \|\mathcal{G}^*_{proj} - \mathcal{G}_R\|_F + \|\mathcal{G}_R - \mathcal{G}\|_F$$

where $\mathcal{G}^*_{proj}$ is defined similarly to $\mathcal{G}^*$, but the vectors used to define it are the result of a Gram-Schmidt orthonormalization process of the vectors used in $\mathcal{G}^*$. $\mathcal{G}_R$ is defined anologously to $\mathcal{G}$, but with the R-approximation matrix T_R instead of T_n.

Inspecting the proof we see that $\|\mathcal{G} - \mathcal{G}_R\|_F \leq \frac{2^{\frac{3}{2}}\|T_n - T_R\|_F}{\Delta}$, were we recognize the residual matrix E_R from the previous Chapter, Section 2.4. In the proof of Cor. 5 we saw that choosing $R = n$ we achieve a parametric rate $\mathcal{O}(\sqrt{\frac{1}{n}})$, that is with probability larger than $1 - \alpha$ we have $\|T_n - T_R\|_F \lesssim_{\alpha,d} \frac{1}{\sqrt{n}}$ (the constant might depend on the dimension d). For the term $\|\mathcal{G}^* - \mathcal{G}^*_{proj}\|_F$ we already have a parametric rate bound. More specifically, we have $\|\mathcal{G}^* - \mathcal{G}^*_{proj}\|_F \lesssim_\alpha \frac{d}{\sqrt{n}}$.

For the term $\|\mathcal{G}^*_{proj} - \mathcal{G}_R\|_F$, we have the bound $\|\mathcal{G}^*_{proj} - \mathcal{G}_R\|_F \leq \frac{\|H\|_F}{\Delta^*}$, where the matrix H was defined in eq (3.12) below. It is easy to see that $\|H\|_F$ is bounded by $\|\Lambda^*_R\|_F \|\Phi_R^T \Phi_R - \mathrm{Id}_R\|_{op}$. From the fact that $s = 0$ and the regularity parameter $\delta > 2$ we have that $\|\Phi_R^T \Phi_R - \mathrm{Id}_R\|_{op} \lesssim_\alpha \frac{1}{\sqrt{n}}$ and the term $\|\Lambda^*_R\|_F$ is bounded by a constant, given that T_W is a Hilbert-Schmidt operator.

In resume we have the following corollary

Corollary 39. *Let W be a regular geometric graphon on $\mathbb{S}^{d-1}$ with regularity parameter $\delta > 2$ and such that $\Delta^* > 0$. Then*

$$\|\hat{\mathcal{G}} - \mathcal{G}^*\|_F = \mathcal{O}_\alpha({\Delta^*}^{-1} n^{-1/2})$$

with probability larger than $1 - \alpha$.

This represent an improvement over Thm. 4. Still this result might be further improved by a better estimation of the quantity $\|E_R\|_{op}$. This would impact specially the value n_0 in the definition of $\mathcal{E}$. Also, a sharper estimation of this term, could lead to improved bounds under weaker assumptions, since in the previous chapter this represented the major bottleneck to remove more regularity hypotheses. We suspect that to make better use of the orthogonality structure, we could use the relative concentration for the eigenvalues converging to λ_1^*. This could help explaining the fact that we have observe an "scaling effect" when the dimension changes in the experiments reported in Fig. 3.1(right).

Another possible strategy is to use the generic chaining results discussed in Chapter 2, but note that in this case the family of functions is not the linear family defined in (2.23) and we predict that sharp bounds will be harder to obtain via this method.

3.10 Some proofs

Proof of Lemma 24. Invoke Theorem 32 with $Y = \hat{T}_n - T_n$, which has independent centered entries conditional to the latent points, to obtain with probability larger than $1 - \alpha$

$$\|\hat{T}_n - T_n\|_{op} \lesssim_\alpha C \max\{\sqrt{\frac{\rho_n}{n}}, \frac{\sqrt{\log n}}{n}\}$$

because $D_0 = \max_{0 \leq i \leq n} \sum_{j=1}^n \Theta_{ij}(1 - \Theta_{ij})$ is $\mathcal{O}(n\rho_n)$, by the definition of Θ. Thus, for n large enough we have

$$\frac{1}{\rho_n}\|\hat{T}_n - T_n\|_{op} \lesssim_\alpha C \max\{\frac{1}{\sqrt{\rho_n n}}, \frac{\sqrt{\log n}}{\rho_n n}\} \leq \frac{(\Delta^*)^2}{2^{\frac{17}{2}}\sqrt{d}}, \tag{3.7}$$

provided that $\frac{\sqrt{\log n}}{\rho_n n} = o(1)$, which holds because we have assume that $\rho_n = \Omega(\log n/n)$.

We use Lemma 37 with $A = \hat{V}\hat{O}$ and $B = V$, where $\hat{O}$ is an orthogonal matrix, and Theorem 29 assuming that the right hand side of (3.6) is smaller than 1, obtaining

$$\begin{aligned} \|\hat{V}\hat{V}^T - VV^T\|_F &\leq 2\|\hat{V}\hat{O} - V\|_F \\ &\leq \frac{2^{\frac{5}{2}} \min\{\sqrt{d}\|T_n - \hat{T}_n\|_{op}, \|T_n - \hat{T}_n\|_F\}}{\Delta} \end{aligned} \tag{3.8}$$

where $\Delta := dist(\{\lambda_{i_1}, \cdots, \lambda_{i_d}\}, \lambda(T_n) \setminus \{\lambda_{i_1}, \cdots, \lambda_{i_d}\})$. Then we have

$$\begin{aligned} \|\hat{V}\hat{V}^T - VV^T\|_F &\leq \frac{2^{\frac{5}{2}} \frac{\sqrt{d}}{\rho_n} \|T_n - \hat{T}_n\|_{op}}{\frac{1}{\rho_n}\Delta} \\ &\lesssim_\alpha \frac{\rho_n (\Delta^*)^2}{2^6 \Delta} \end{aligned} \tag{3.9}$$

where the last inequality holds under $\mathcal{E}$ and with confidence at least $1-\alpha$. From Theorem 33 we have that, when n is large enough

$$\delta_2\Big(\lambda(\frac{1}{\rho_n}T_n), \lambda(T_W)\Big) \lesssim_\alpha C\Big(\frac{\log n}{n}\Big)^{\frac{\delta}{2\delta+d-1}} \leq \frac{\Delta^*}{8}\,, \tag{3.10}$$

where the last inequality is valid under $\mathcal{E}$. This and (3.7) ensure that exist n_0 that depends on Δ^* and α such that $\mathbb{P}(\mathcal{E}) \geq \alpha/2$. □

Proof sketch of Prop.25. When $\Delta^* > 0$, we remark that λ_1^* is the only eigenvalue of T_W with multiplicity $d_1 = d$, the others eigenvalues (except for λ_0^*) having multiplicity strictly greater than d. Now, using (3.10) we deduce that, under $\mathcal{E}$, there exists a unique set $\frac{1}{\rho_n}\lambda_{i_1}, \frac{1}{\rho_n}\lambda_{i_2}, \dots, \frac{1}{\rho_n}\lambda_{i_d}$ of d eigenvalues of T_n that can be separated from the other eigenvalues by a distance at least $3\Delta^*/4$, namely the triangular inequality gives

$$\frac{\Delta}{\rho_n} \geq \frac{3\Delta^*}{4}\,. \tag{3.11}$$

Furthermore, using (3.9) we get that there exists eigenvalues $\hat{\lambda}_{i_1}, \hat{\lambda}_{i_2}, \dots, \hat{\lambda}_{i_d}$ and eigenvectors $\hat{V} \in \mathbb{R}^{n\times d}$ of $\hat{T}_n$ such that $\|\hat{V}\hat{V}^T - VV^T\|_F \leq \Delta^*/48$. We define $\Lambda_1 := \{\hat{\lambda}_{i_1}, \cdots, \hat{\lambda}_{i_d}\}$. By Hoffman-Wielandt inequality [Bathia 1997, Thm.VI.4.1], it holds

$$\Big(\sum_{k=1}^d (\hat{\lambda}_k^{\text{sort}} - \lambda_k^{\text{sort}})^2\Big)^{1/2} \leq \|\hat{V}\hat{V}^T - VV^T\|_F \leq \Delta^*/8\,,$$

By triangular inequality, we deduce that

$$\hat{\Delta} := dist(\Lambda_1, \lambda(\frac{1}{\rho_n}\hat{T}_n) \setminus \Lambda_1) \geq \frac{\Delta^*}{2}\,,$$

namely $\hat{\lambda}_{i_1}, \hat{\lambda}_{i_2}, \dots, \hat{\lambda}_{i_d}$ is a set of d eigenvalues at distance at least $\Delta^*/2$ from the other eigenvalues of $\hat{T}_n$.

This analysis can be also done for the other eigenvalues, which shows that, on the event $\mathcal{E}$, Algorithm 1 returns $\hat{\mathcal{G}} = (1/c_1)\hat{V}\hat{V}^T$ composed by the eigenvectors corresponding to the eigenvalues of the aforementioned cluster of d eigenvalues. □

Proof sketch of Thm. 26. We have by (3.8) that

$$\|\hat{\mathcal{G}} - \mathcal{G}\|_F = \frac{1}{c_1}\|\hat{V}\hat{V}^T - VV^T\|_F \lesssim_\alpha C\frac{(d-2)}{\sqrt{dn}},$$

whenever n is large enough and $\Delta^* > 0$, where C may depend on W. In the last inequality we used that $c_1 = d/(d-2)$.

We are left to control $\|\mathcal{G}^* - \mathcal{G}\|_F$. We have by triangle inequality

$$\|\mathcal{G}^* - \mathcal{G}\|_F \leq \|\mathcal{G}^* - \mathcal{G}^*_{proj}\|_F + \|\mathcal{G}^*_{proj} - \mathcal{G}_R\|_F + \|\mathcal{G}_R - \mathcal{G}\|_F$$

where $\mathcal{G}^*_{proj}$ is the projection matrix for the column span of the matrix V^* and $\mathcal{G}_R$ is the Gram matrix for the eigenvectors of the approximation matrix $T_R = \left(\frac{1}{n}W_R(X_i, X_j)\right)_{i,j}$.

We use Theorem 29 to obtain

$$\|\mathcal{G} - \mathcal{G}_R\|_F \leq \frac{2^{\frac{3}{2}}\|T_n - T_R\|_F}{\Delta} \leq C\frac{(n/\log n)^{-\delta/(2\delta+d-1)}}{\Delta}$$

where we choose $R = \mathcal{O}((n/\log n)^{\frac{1}{2\delta+d-1}})$. We use Lemma 38 with $B = V^*$ and Theorem 34 obtaining

$$\|\mathcal{G}^* - \mathcal{G}^*_{proj}\|_F \leq \|\mathrm{Id}_d - {V^*}^T V^*\|_F \lesssim_\alpha \frac{d}{\sqrt{n}}$$

We have, see [Bathia 1997, p.202]

$$\|\mathcal{G}^*_{proj} - \mathcal{G}_R\|_F = 2\|\mathcal{G}^*_{proj}{\mathcal{G}_R}^{\perp}\|_F$$

We use Theorem 31, with $E = \mathcal{G}^*_{proj}$, $F = \mathcal{G}^{\perp}_R$, $B = T_R$ and $A = T_R + H$, where

$$H := \tilde{\Phi}_R \Lambda^*_R \tilde{\Phi}^T_R - \Phi_R \Lambda^*_R \Phi^T_R \tag{3.12}$$

the matrix $\tilde{\Phi}_R$ has column $\tilde{\Phi}_k$ obtained from Φ_k by a Gram-Schmidt orthonormalization process. By Theorem 31 we have

$$\|\mathcal{G}^*_{proj}\mathcal{G}^{\perp}_R\|_F \leq \frac{\|A - B\|_F}{\Delta^*} = \frac{\|H\|_F}{\Delta^*}$$

To bound $\|H\|_F$, we use Ostrowskii's inequality Cor. 18 and [De Castro 2020, Lem.12] to obtain

$$\|\mathcal{G}^*_{proj}\mathcal{G}^{\perp}_R\|_F \lesssim_\alpha \frac{1}{\Delta^*}\left(\frac{\log n}{n}\right)^{\frac{\delta}{2\delta+d-1}}$$

which implies that

$$\|\mathcal{G}^*_{proj} - \mathcal{G}_R\|_F \lesssim_\alpha \frac{1}{\Delta^*}\left(\frac{n}{\log n}\right)^{\frac{-\delta}{2\delta+d-1}}$$

We conclude by collecting terms.

□

Proof of Lemma 27 . The lemma follows from Proposition 25. Indeed, on the event $\mathcal{E}$ there exists only one set Λ_1 of eigenvalues of $\hat{T}_n$ with cardinality d , whose distance to the rest of the spectrum is larger that Δ^* and its diameter is smaller that Δ^*. When sorting the eigenvalues of $\hat{T}_n$ in decreasing order, those belonging to Λ_1 will appear in consecutive order. The lemma follows from this observation and from the fact $\text{Gap}_1(\hat{T}_n; i_{n-d-1}, \cdots, i_{n-1}) = \text{left}(n-d-1)$. □

Proof of 32. By [Bandeira 2016, Rmk.3.13] we have the tail concentration bound (taking their ε equal to 1/2)

$$\mathbb{P}\Big(\|Y\|_{op} \geq 3\sqrt{2D_0} + \max_{ij} \|Y_{ij}\|_\infty C_0 \sqrt{\log n/\alpha}\Big) \leq \alpha$$

the result follows, because $\max_{ij} \|Y_{ij}\|_\infty \leq 1$. □

Proof of Proposition 36 . By Courant-Fisher min-max principle we have

$$\lambda_0^* = \max_{f \in L^2([-1,1])} \frac{\langle T_W f, f\rangle}{\langle f, f\rangle}$$

In particular, if we take the function $\mathbb{1}(x) := 1$ for $x \in [-1,1]$ we have

$$\begin{aligned}\lambda_0^* &\geq \frac{\langle T_W \mathbb{1}, \mathbb{1}\rangle}{\langle \mathbb{1}, \mathbb{1}\rangle} \\ &= \frac{\int_{\mathbb{S}^{d-1}} W(x,y) d\sigma(x) d\sigma(y)}{\int_{\mathbb{S}^{d-1}} d\sigma(y)} \\ &= d_W\end{aligned}$$

the last follows form the definition of d_W and the fact that σ is a probability measure on the sphere. On the other hand, if f_0 is an eigenfunction associated with λ_0 we can choose x^* such that $f_0(x^*) \geq f_0(x)$ for $x \in [-1,1]$. Without loss of generality, assume that $f_0(x^*) \neq 0$. So

$$\begin{aligned}\lambda_0^* &= \frac{T_W f_0(x^*)}{f_0(x^*)} \\ &= \int_{\mathbb{S}^{d-1}} W(x^*, y) \frac{f_0(y)}{f_0(x^*)} d\sigma(y) \\ &\leq \int_{\mathbb{S}^{d-1}} W(x^*, y) d\sigma(y) \\ &= d_W(x^*)\end{aligned}$$

which finish the proof □

Chapter 4

Random Geometric Graphs on the Euclidean Ball

Contents

4.1 Introduction

In this chapter we introduce the model of the random geometric graphs on the Euclidean ball $\mathbb{B}^d = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$, where we assume $d \geq 3$, which can be considered as an extension of the RGG model on the sphere discussed in Chapter 3, since it is defined by a dot product kernel and the graphon formalism. Although the family of RGG graphs on $\mathbb{B}^d$ we study in this chapter has an intersection with the family of random dot product graphs (RDPG) [Athreya 2018], the model we describe here has not been previously studied with detail in the literature. Consequently, the purpose of this chapter is twofold: to show that, similar to the case of RGG model on $\mathbb{S}^{d-1}$, it is well suited to latent structure recovery problems and to highlight some of the properties that make it appealing for the modeling of more complex real networks.

As in the previous chapter we will consider randomly placed latent points $\{X_i\}_{1\leq i\leq n}$, whose connection probability will depend on a one dimensional link function evaluated on the inner products of the form $\langle X_i, X_j \rangle$. One key difference with the RGG model on $\mathbb{S}^{d-1}$ is that on the ball the probability of obtaining a graph with a more heterogenous degree sequence is frequently higher than in the sphere. This is due to the fact the points $\{X_i\}_{1 \leq i \leq n}$ have different norms and their inner product depend not only on the angles, but also on the norms. Take for instance the classical RGG model with link function $\mathbb{1}_{\langle X_i, X_j \rangle \geq \tau}$. A latent point with large norm will be more likely to be connected than one with smaller norm. This creates a heterogeneity in the node degree sequence that is not present on the RGG model on $\mathbb{S}^{d-1}$. The fact that some nodes are more connected than others represent a feature that many real network share (see [Barrat 2004] for example) and, consequently, the RGG model on the ball will offer more flexibility in terms of modelling.

Even if the both models (the one on the sphere and the one on the ball) can generate very different graphs, they do have formal similarities that will allow us to extend the analysis from the previous chapter. Indeed, in the ball there is also a fixed basis of orthogonal polynomials that plays the role that the spherical harmonics had in the previous chapter. The harmonic analysis on $\mathbb{B}^d$ provides an explicit representation for the reproducing kernel for the space of polynomials of each degree, which gives formulas for the spectral expansion which share similarities with the spherical case. In particular, the linear polynomial contains the inner product information, which can be recovered from the eigenfunctions. On the other hand, our analysis will be valid not only for the uniform distribution on $\mathbb{B}^d$, but also for a parametric family of spherically symmetric distributions related to the beta distribution. Even if this mainly a technical assumption that we inherit from the harmonic analysis on the ball [Dai 2013, Chap.11], it is also an added feature with respect to the spherical model that gives more flexibility to the model and makes it capable to express more complex networks.

We discuss the problem of recovering latent information on this model, mainly under a spectral gap assumption. We discuss the possibility of estimation of the latent norm from the observed adjacency matrix, in the threshold graphon model. We also discuss the estimation of the latent distance, extending the approach developed for the sphere in Chapter 3. Some related problems, involving the recovery of latent structures, have recently been studied in [Athreya 2020], from the spectral point of view, but on the RDPG model and with distributional assumptions of the latent points and ambient spaces different from the ones we consider here.

4.2 RGG on the ball

In this section we define the RGG model on the Euclidean ball. We will introduce a set of measures that will be admissible in our model. Part of the material presented here is classic in the context of harmonic analysis on $\mathbb{B}^d$ [Dai 2013, Chap.11], including the geometric formulas on Euclidean spaces with measures using Jacobi weights.

We define $\mathcal{F} = \{F_\nu\}_{\nu > -1/2}$ the parametric family of distributions on $\mathbb{B}^d$ with densities, with respect to the Lebesgue measure, given by

$$dF_\nu(x) = C_\nu(1 - \|x\|^2)^{\nu - \frac{1}{2}}$$

where

$$C_\nu = \int_{\mathbb{B}^d} (1 - \|x\|^2)^{\nu - \frac{1}{2}} dx$$

By definition each distribution on the family $\mathcal{F}$ is spherically symmetric [Kelker 1970] which in the case of distributions which are absolutely continuous with respect to the Lebesgue measure is given by the fact that for every $x \in \mathbb{B}^d$ the density $dF_\nu(x)$ depends on the norm of x only. Observe that for $\nu = \frac{1}{2}$ the distribution F_ν is equal to the uniform distribution on $\mathbb{B}^d$. If X is $\mathbb{B}^d$-valued random variable with law F_ν, then the following stochastic representation holds

$$X = RU$$

where R is a real valued random variable distributed as $\|X\|$ and U is and independent uniform random direction, that is, its law is given by the uniform measure on the sphere σ, defined in the previous chapter. The variable U has the same distribution that $X/\|X\|$. The following lemma describe the distribution of $\|X\|$

Lemma 40. *If X is a $\mathbb{B}^d$-valued random variable distributed according to F_ν, then $\|X\|^2$ follows a distribution $Beta(\frac{d}{2}, \nu + \frac{1}{2})$.*

The following procedure characterizes th RGG model on the ball. First, we sample i.i.d points $\{X_i\}_{1 \leq i \leq n}$ according to $F_\nu \in \mathcal{F}$, for some $\nu > 0$. Then, conditional to those points we sample the adjacency matrix A_{ij} such that for $i < j$

$$\mathbb{P}(A_{ij} = 1) = f(\langle X_i, X_j \rangle)$$

for some function f defined on $[-1, 1]$. The entries A_{ij} for $i > j$ are defined by symmetry.

Examples of this model are the Erdös-Rényi model, where $f(t) = p$ for $p \in [0, 1]$ and threshold or proximity graphon $f(t) = \mathbb{1}_{t \geq \tau}$ for $\tau \geq 0$. Notice that in general, the class of proximity graph is a subclass of the intersection graphs where each node has associated a set and an arc exist between two nodes whenever the two associated sets intersect. In the case of proximity graphons on $\mathbb{S}^{d-1}$ the set associated with each node is the boundary of a spherical cap (a hypersurface in $\mathbb{R}^d$) around that node and with height $1 - \tau$. In the case of the proximity graphon on the ball the associated set is a spherical cap, but the heigh is not constant and depend on the node in consideration and also on τ.

Let W be a graphon defined on $\mathbb{B}^d$ with measure μ, the (normalized) degree function is defined as follows [Lovasz 2012][Chap. 7]

$$d_W(x) = \int_{\mathbb{B}^d} W(x, y) d\mu(y)$$

In the case of a geometric graphon $W(x,y) = f(\langle x,y\rangle)$, it is easy to see that the $d_W(x) = d_W(x')$ when $\|x\| = \|x'\|$. Observe that in the case of the Erdös-Rényi model we have a constant degree function, for any measure μ. For the threshold function $W_g(x,y) = \mathbb{1}_{\langle x,y\rangle \geq \tau}$, we consider the measure $\mu = F_\nu$, for some ν [1]. Then we have for the degree function

$$\begin{aligned} d_{W_g}(x) &= \int_{\mathbb{B}^d} \mathbb{1}_{\langle x,y\rangle \geq \tau} dF_\nu(x) \\ &= F_\nu\Big(\mathrm{Sc}\left(x, 1 - \frac{\tau}{\|x\| \vee \tau}\right)\Big) \end{aligned}$$

where $\mathrm{Sc}(x,h)$ represents the spherical cap on $x/\|x\|$ with heigh h, that is

$$\mathrm{Sc}(x,h) := \{y \in \mathbb{B}^d : \langle y, x/\|x\|\rangle \geq 1 - h\}$$

Fix $X_i \in \mathbb{B}^d$, then the probability that X_j is connected to X_i for $j \neq i$ is

$$\mathbb{P}(A_{ij} = 1) = F_\nu\Big(\mathrm{Sc}\left(X_i, 1 - \frac{\tau}{\|X_i\| \vee \tau}\right)\Big)$$

Note that if $\|X_i\| \geq \tau$, then the spherical cap in the previous formula reduce to a point and, therefore, has measure zero. In other words, the points with $\|X\| \leq \tau$ are disconnected from the rest of the graph. For a fixed i the random variables A_{ij} for $1 \leq j \leq n$ are independent. Denote $d_G(X_i) := \sum_{j\neq i} A_{ij}$ the degree of X_i in the random graph. Observe that the random variable $d_G(X_i)$, conditional to X_i, follows a distribution $\mathit{Binomial}(n-1, d_W(X_i))$, thus

$$\mathbb{E}\Big(\frac{d_G(X_i)}{n-1}\Big) = d_{W_g}(X_i) = F_\nu(\mathrm{Sc}(X_i, 1 - \tau/\|X_i\|)) \tag{4.1}$$

From the fact that F_ν is spherically symmetric, we deduce that

$$F_\nu(\mathrm{Sc}(x, 1 - \tau/\|x\|) = F_\nu(\mathrm{Sc}(x', 1 - \tau/\|x'\|)$$

if $x = \|x'\|$. In other words, the previous quantity depends only on the norm of x. From standard concentration inequalities, such as the Hoeffding inequality, and from (4.1) we have the following lemma

Lemma 41. *Fix $X_i \in \mathbb{B}^d$. Given $\varepsilon > 0$ we have that*

$$\mathbb{P}\big(\big|\frac{d_G(X_i)}{n-1} - d_{W_g}(X_i)\big| > \varepsilon\big) \leq 2\exp -2\varepsilon^2 n$$

In words, the random variable $\frac{d_G(X_i)}{n-1}$ is highly concentrated, for n large, around its mean $d_{W_g}(X_i)$. Intuitively speaking, if we fix a node X_i in the graph its degree will reveal information about the latent norm $\|X_i\|$. We examine this closer in Section 4.3.

[1] With some abuse of notation we use F_ν for the distribution function and the measure.

To highlight the differences between the RGG models in the sphere and in the ball, let consider the function W_g with $\tau > 0$. We have the same connection rule, but different underlying measures. In the case of the sphere the measure is the uniform σ (the only rotation-invariant Borel measure on $\mathbb{S}^{d-1}$) and in the case of the ball, let us consider the uniform measure, that is, $F_{1/2}$. Let call d_{sphere} and d_{ball} the degree function for the models in the sphere and in the ball. Using standard formulas for the area and volume of spherical caps in $\mathbb{R}^d$, we see that the function $d_{sphere}(\cdot)$ is constant and equal to

$$d_{sphere}(x) = \sigma(\mathrm{Sc}(N, 1-\tau)) = \frac{1}{2} I_{1-\tau^2}\left(\frac{d-1}{2}, \frac{1}{2}\right) =: d_{s,\tau}$$

where $I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1}$ is the regularized incomplete Gamma function and $N = (1, \cdots, 0) \in \mathbb{R}^d$ is the north pole on $\mathbb{S}^{d-1}$. On the other hand, we have

$$d_{ball}(x) = F_{1/2}(\mathrm{Sc}(N, 1 - \frac{\tau}{\|x\| \vee \tau})) = \frac{1}{2} I_{1-(\frac{\tau}{\|x\| \vee \tau})^2}\left(\frac{d+1}{2}, \frac{1}{2}\right)$$

Notice that, using the properties of the incomplete Beta function, we have for $x \in \mathbb{S}^{d-1}$

$$d_{ball}(x) = d_{sphere}(x) - \frac{1}{(d-1)B(\frac{d-1}{2}, \frac{1}{2})} \tau(1-\tau^2)^{\frac{d-1}{2}}$$

From the previous we see that when $\tau = 0$ we have $d_{ball}(x) = d_{sphere}(x)$ for $x \in \mathbb{S}^{d-1}$. Indeed, in the case of $\tau = 0$ we have $d_{ball}(x) = d_{ball}(\cdot)$, because $I_1(a,b) = 1$ for any $0 \leq a, b \leq 1$. Consequently, the model in the sphere and the model in the ball, for $\tau = 0$, define the same distribution over graphs.

For the case $\tau > 0$ we have for $\|x\| \leq \tau$ we have $d_{ball}(x) = 0$. Thus $d_{ball}(x) \in [0, \tilde{d}_{s,\tau}]$ for any $x \in \mathbb{B}^d$, where

$$\tilde{d}_{s,\tau} = d_{s,\tau} - \frac{1}{(d-1)B(\frac{d-1}{2}, \frac{1}{2})} \tau(1-\tau^2)^{\frac{d-1}{2}}$$

Given the concentration result Lemma 41, we expect that for τ bounded away from 0 to observe a different degree distribution in both models, assuming the same τ and the same sample size. On one hand, in the spherical model, we expect the degree of every node in the graph to be very similar for n large. In the case of the model in $\mathbb{B}^d$, the degree sequence will be more heterogeneous. We discuss this further in the next section.

Random graph models are typically compared fixing the mean degree of the resulting graph, that is, fixing $\bar{d}_G = \frac{1}{n} \sum_{i=1}^n d_G(X_i)$. This is common in the context of testing, where we need to decide between two models given a single observation of the graph. See for example [Arias-Castro 2014, Arias-Castro 2015] for testing for the SBM and [Bubeck 2016] for the testing problem on angular RGG's on $\mathbb{S}^{d-1}$. If we fix $\tau_1 \geq 0$, and define $W_1(x,y) = \mathbb{1}_{\langle x,y\rangle \geq \tau_1}$, in the sphere. We can find τ_2, which defines $W_2(x,y) = \mathbb{1}_{\langle x,y\rangle \geq \tau_2}$, in the ball, such that both models have the same mean degree function, that is $d_{W_1}(x) = \int_B d_{W_2}(x) dF_{1/2}(x)$. Also note that the concentration study for the random variable $\bar{d}_G$ can be carried out in this case and is equivalent to the the concentration of the matrix norm $\|A\|_{1,1} = \sum_{i,j} |A_{ij}|$ for the adjacency matrix. We do not pursue such a study here.

We illustrate the previous point by simulations. We consider W_1 and W_2 as previously defined with $\tau_1 = 1/2$ and $\tau_2 = 1/4$. We sample graphs 25 graphs from both models, each of size $n = 1000$ and compute the minimum degree, the maximum degree and the mean degree. In Figure 4.1 we show an example of one of the instances obtained for each graph. In Table 4.1 we report the mean over the 25 simulations for the basic statistics regarding the degree sequence.

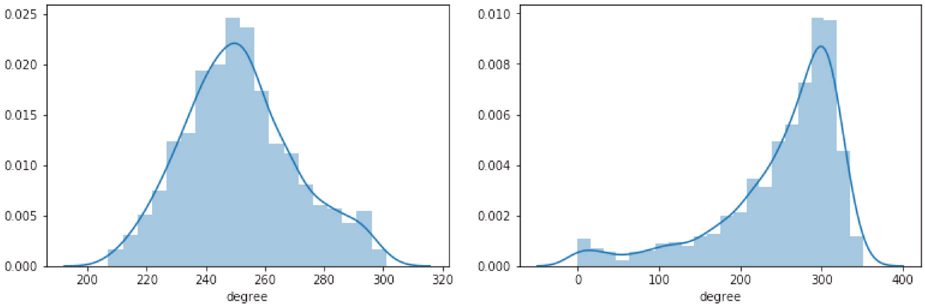

Figure 4.1: Degree histogram for W_1 (left) and W_2 (right).

Model	Min.deg	Mean deg	Max.deg
W_1	199	250.672	282
W_2	0	249.682	342

Table 4.1: Table with the mean of the *min, max and mean degree* statistics for each model.

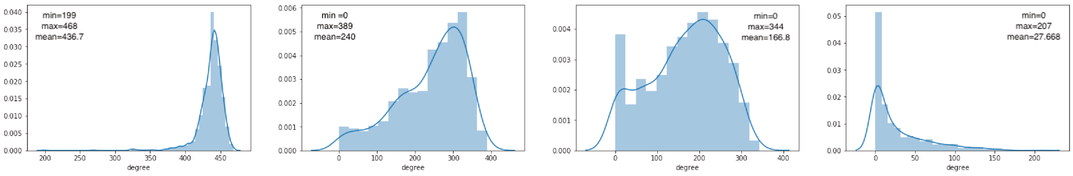

Figure 4.2: Degree histogram for W_g with $\tau = 0.1$ and $\nu = -0.3$(left),$\nu = 3.5$,$\nu = 5.5$ and $\nu = 15.5$(right).

In Figure 4.2 we show the degree histogram for graphs sampled from the graphon W_g with fixed $\tau = 0.1$ and for $\nu = -0.3, 3.5, 5.5, 15.5$. The purpose is to illustrate the variety of different shapes that the degree distribution can have on this model. We repeat this exercise, but fixing $\nu = 5.5$ this time and changing the value of τ, which is shown in Figure 4.3. Observe that as τ grows, not only it changes the shape of the degree distribution but also the graph become sparser.

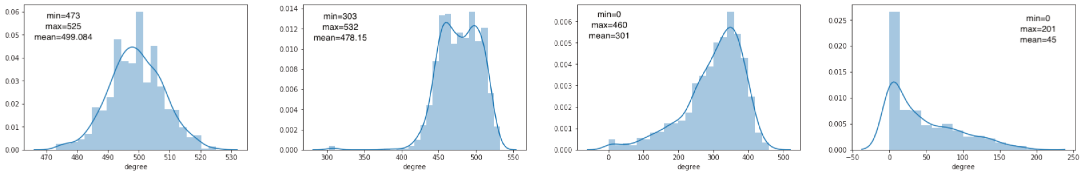

Figure 4.3: Degree histogram for W_g with $\nu = 5.5$ and $\tau = 0$(left),$\tau = 0.005$,$\tau = 0.05$ and $\tau = 0.2$(right).

4.2.1 The sough after power law distribution

In previous section we saw that, in terms of observed degree distribution, the RGG model on the ball is more flexible that the one on the sphere. Given that the heterogeneity in the degree sequence is a characteristic observed in many real world networks [Barrat 2004], having a more flexible model at hand would be useful for modeling purposes. From (4.1) we see that the possible values for $d_W(x)$ for a threshold graphon W are of the form $F_\nu(\mathrm{Sc}(N, 1 - \frac{\tau}{\|x\| \vee \tau}))$. However useful the previous characterization might be, it has the problem of not being very explicit and as it is written do not match any of the typical degree distributions that are frequent in the network literature. In particular, many real networks exhibit a power law degree distribution [Clauset 2009, Mitzenmacher 2003], meaning that the number of nodes with (unnormalized) degree k is proportional to $k^{-\gamma}$ with $\gamma > 0$. This opens the question: is the power law included between the possible degree distributions in the RGG model on the ball? or at least, is it possible to have an approximative version of it?

We will first explore the degree for the case of the graphon W_g, whose degree function is given by (4.1). More specifically, we have

Proposition 42. *For the threshold graphon $W = W_g$ for $\tau \geq 0$ and $\{X_i\}_{1 \leq i \leq n} \sim F_\nu$ for $\nu > -1/2$, we have for any $1 \leq i \leq n$*

$$\mathbb{P}\big(d_W(t_1 N) \leq d_W(X_i) \leq d_W(t_2 N)\big) = F_{Beta}(t_2^2) - F_{Beta}(t_1^2)$$

where $0 < \tau < t_1 < t_2$ and $F_{Beta}(\cdot)$ is the cumulative distribution function of Beta ($\frac{d}{2}$, $\nu +$ $\frac{1}{2}$). In addition, we have that

$$\mathbb{P}(d_W(X_i) = 0) = F_{Beta}(\tau)$$

Notice that Prop. 42 characterizes the distribution of the degree function. However, its application depend on the computation of $d_W(tN)$ for different values of $t \in [0, 1]$. This corresponds to compute the volume against the measure F_ν for spherical cap $\mathrm{Sc}(N, 1 - \frac{\tau}{t \vee \tau})$. In that regard, we have the following integration lemma

Lemma 43. *For $\tau \geq 0$ we have*

$$d_W(tN) = \frac{1}{2} I_{1 - \left(\frac{\tau}{t \vee \tau}\right)^2}\big(\nu + \frac{d}{2}, \frac{1}{2}\big)$$

It is clear that the previous lemma and Prop. 42 characterize the distribution function of the degree function $d_W(x)$. More specifically, the distribution function $F_{d_W}(\cdot)$ for $d_W(X)$, when $X \sim F_\nu$ is given by

$$F_{d_W}(t) = \begin{cases} I_{\tau^2}(\frac{d}{2}, \nu + \frac{1}{2}) \text{ if } t = 0 \\ I_{g(t)}(\frac{d}{2}, \nu + \frac{1}{2}) \text{ if } t > 0 \end{cases}$$

where $g(t) = \frac{\tau^2}{1 - I^{-1}(2t; \frac{d}{2} + \nu, \frac{1}{2})}$.

In order to find an analog to the power law distribution we will need to compute explicitly the degree function (while finding an explicit bound would suffice) and the previous expression has the drawback of depending on the complicated regularized incomplete Beta function and its inverse (which cannot be expressed using standard functions). We will, nonetheless, use the previous expression in the next section when we study the latent norm recovery.

To continue the search for a similar distribution to the power law, we study the degree function of the following RGG on $\mathbb{B}^d$, defined by the connection function:

$$f(t) = \begin{cases} \frac{\alpha}{|t|^2} \wedge 1 & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

where $\alpha \in (0, 1)$ is a "resolution" parameter. For the latter we mean that if $x \in \mathbb{B}^d(0, \sqrt{\alpha})$ then for all $y \in \mathbb{B}^d$ we have $f(\langle x, y \rangle) = 1$. That is, any point located in the ball centered in 0 with radius α will connect with every other point in $\mathbb{B}^d$. This is the inverse of what happens in the threshold graphon. In terms of the degree function, this means that $d_f(x) = 1$ for $\|x\|^2 \leq \alpha$. By defintion we have

$$d_f(x) = \int_{\mathbb{B}^d} \frac{\alpha}{|\langle x, y \rangle|^2} \wedge 1 dF_\nu(y)$$

By rotational symmetry (we can think of x being $x = (x_1, 0, \cdots, 0) = x_1 N$) we have

$$\begin{aligned} d_f(x) &= d_f(x_1 N) \\ &= \int_{\mathbb{B}^d} \frac{\alpha}{x_1^2 y_1^2} \wedge 1 dF_\nu(y) \\ &= \int_{\mathbb{B}^d \setminus \mathbb{B}^d(0,\sqrt{\alpha})} \frac{\alpha}{x_1^2 y_1^2} \wedge 1 dF_\nu(y) + \int_{\mathbb{B}^d(0,\sqrt{\alpha})} dF_\nu(y) \end{aligned}$$

Given that the summand $\int_{\mathbb{B}^d(0,\sqrt{\alpha})} dF_\nu(y)$ is common to every $x \in \mathbb{B}^d$ we will subtract it. Intuitively speaking, there will be nodes that will be connected will almost every node in the graph, which increase the minimum degree. Since we already know that the nodes such that $\|x\|^2 \leq \alpha$ are connected with every other node, we concentrate in the case $\|x\|^2 > \alpha$. This motivates the definition

$$\tilde{d}_f(x) := \frac{\int_{\mathbb{B}^d \setminus \mathbb{B}^d(0,\sqrt{\alpha})} \frac{\alpha}{x_1^2 y_1^2} dF_\nu(y)}{\int_{\mathbb{B}^d \setminus \mathbb{B}^d(0,\sqrt{\alpha})} \frac{1}{y_1^2} dF_\nu(y)}$$

for $\|x\|^2 > \alpha$. Let take $k, n' \in \mathbb{N}$ such that $k < n' \leq n$. By definition

$$\tilde{d}_f(\sqrt{\frac{n'\alpha}{k}}N) = \frac{\int_{\mathbb{B}^d \setminus \mathbb{B}^d(0,\sqrt{\alpha})} \frac{k}{n' y_1^2} dF_\nu(y)}{\int_{\mathbb{B}^d \setminus \mathbb{B}^d(0,\sqrt{\alpha})} \frac{1}{y_1^2} dF_\nu(y)} = \frac{k}{n'}$$

Thus, given that X is distributed according to F_ν

$$\mathbb{P}(\frac{k-1}{n'} \leq d_f(X) \leq \frac{k}{n'}) = \mathbb{P}\big(X \in \mathbb{B}^d(0, \sqrt{\frac{n'\alpha}{k-1}}) \setminus \mathbb{B}^d(0, \sqrt{\frac{n'\alpha}{k}})\big)$$
$$= F_\nu\Big(\mathbb{B}^d(0, \sqrt{\frac{1}{k-1}}) \setminus \mathbb{B}^d(0, \sqrt{\frac{1}{k}})\Big)$$
$$= I_{\frac{n'\alpha}{k-1}}(\frac{d}{2}, \nu + \frac{1}{2}) - I_{\frac{n'\alpha}{k}}(\frac{d}{2}, \nu + \frac{1}{2})$$

Since $I_x(\cdot, \cdot)$ is an increasing function we see that

$$\Delta_k \frac{d}{dx} I(\frac{n'\alpha}{k-1}) \leq I(\frac{n'\alpha}{k-1}) - I(\frac{n'\alpha}{k}) \leq \Delta_k \frac{d}{dx} I(\frac{n'\alpha}{k})$$

where $I(x) = I_x(\frac{d}{2}, \nu + \frac{1}{2})$ and $\Delta_k = \frac{n'\alpha}{k-1} - \frac{n'\alpha}{k}$. It follows from the definition that $\frac{d}{dx} I_x(a, b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$ and consequently

$$c_{\nu,\alpha}(\frac{k-1}{n'})^{-\frac{d}{2}-1} \leq I(\frac{1}{k-1}) - I(\frac{1}{k}) \leq C_{\nu,\alpha}(\frac{k}{n'})^{-\frac{d}{2}-1}$$

where $c_{\nu,\alpha}$,$C_{\nu,\alpha}$ are constants that depend on ν and α. Thus picking $d = 3$, for example, we have that

$$\mathbb{P}(\frac{k-1}{n'} \leq \tilde{d}_f(X) \leq \frac{k}{n'}) \propto (k/n')^{-2.5}$$

which can see as a similar distribution to a power law[2]. To make this point clearer, take for example n' of order $\mathcal{O}(n)$.The previous can be interpreted as the proportion of nodes of degree k, for k large, follows a power law function, after shifting. The exponent -2.5 has been frequently reported in the literature for real networks [Clauset 2009]. We see that that changing the exponent α in the definition of $f(t)$ and the changing the dimension of the sphere, we can fine-tune the power law exponent parameter.

Remark 8. *There is no formal definition of a power law or shifted power law distribution in the context of graphons. However, what we did here is close to what is described in [Borgs 2019], where the theory of unbounded L^p-graphons is developed. This fact is also present in our definition of $\tilde{d}_f(\cdot)$ which we could be have defined with the term $\int_{\mathbb{B}^d \setminus \mathbb{B}^d(0,\sqrt{\alpha})} \frac{\alpha}{x_1^2 y_1^2} \wedge 1 dF_\nu(y)$ in the numerator, but this makes less evident the analogy with the discrete power law.*

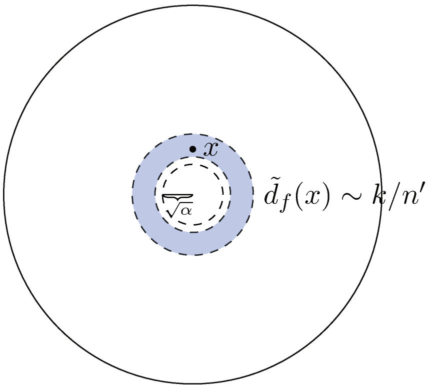

Figure 4.4: A point x inside the annulus between the circles $\sqrt{\frac{n'\alpha}{k-1}}$ and $\sqrt{\frac{n'\alpha}{k}}$ will have degree function satisfying $\tilde{d}_f(x) \sim k/n'$. The fraction of points with degree k/n' is the measure of the annulus.

We end this section by showing simulations for the the model with connection function $f(t)$, to illustrate empirically the behavior of the degree profile under this model. In Figure 4.5 we show a single simulation of the graph of size 3000 with connection function $f(\cdot)$ and parameter $\alpha = 1/1000$, under the measure $F_{1/2}(\cdot)$. We observe the presence of nodes with very high degree (or *large hubs*) which is often common in real world networks and scale-free networks. We include a $\log - \log$ plot for the histogram for the nodes with degrees over 300, to better observe the exponential decay. The resulting shape, first close to a line and then oscillatory (Figure 4.5(right)), has been reported in real world networks, where it is suggested as evidence for a power law distribution of the degrees [Clauset 2009].

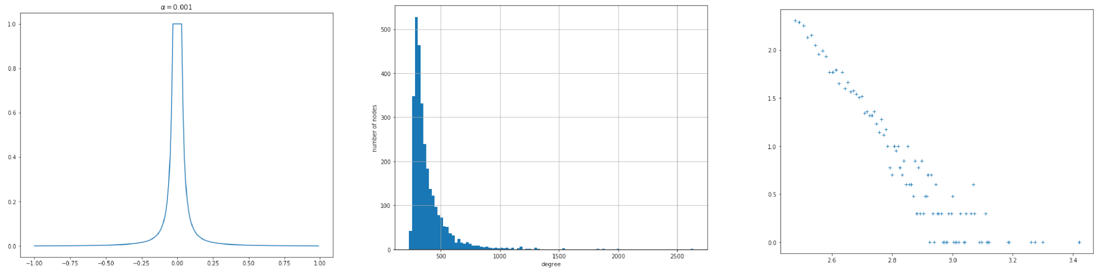

Figure 4.5: From left to right: we plot the function $f(\cdot)$ for $\alpha = 1/1000$. In the center, we show the histogram for this f. In the right, we show a $\log - \log$ plot of the same histogram, but only for the values with degree larger than 300.

We repeat this exercise in Figure 4.6, for different values of α which produces changes in the distribution. We opt to include the $\log - \log$ plot for nodes with degree larger than 300 for comparison purposes. This shows the shifted power law shape of the degree distribution. More node connectivity can also be achieved by

[2] A random graph model has power law if the number of nodes with (unnormalized) degree k is proportional to $k^{-\gamma}$ with $\gamma > 0$.

changing the measure under which we simulate. We show one example in the image at the bottom in Figure 4.6, which was generated with $F_{3/2}$. Indeed, a measure that allocates more mass at the center, will have larger connectivity with this model. This serves to illustrate the flexibility and expressiveness of this model.

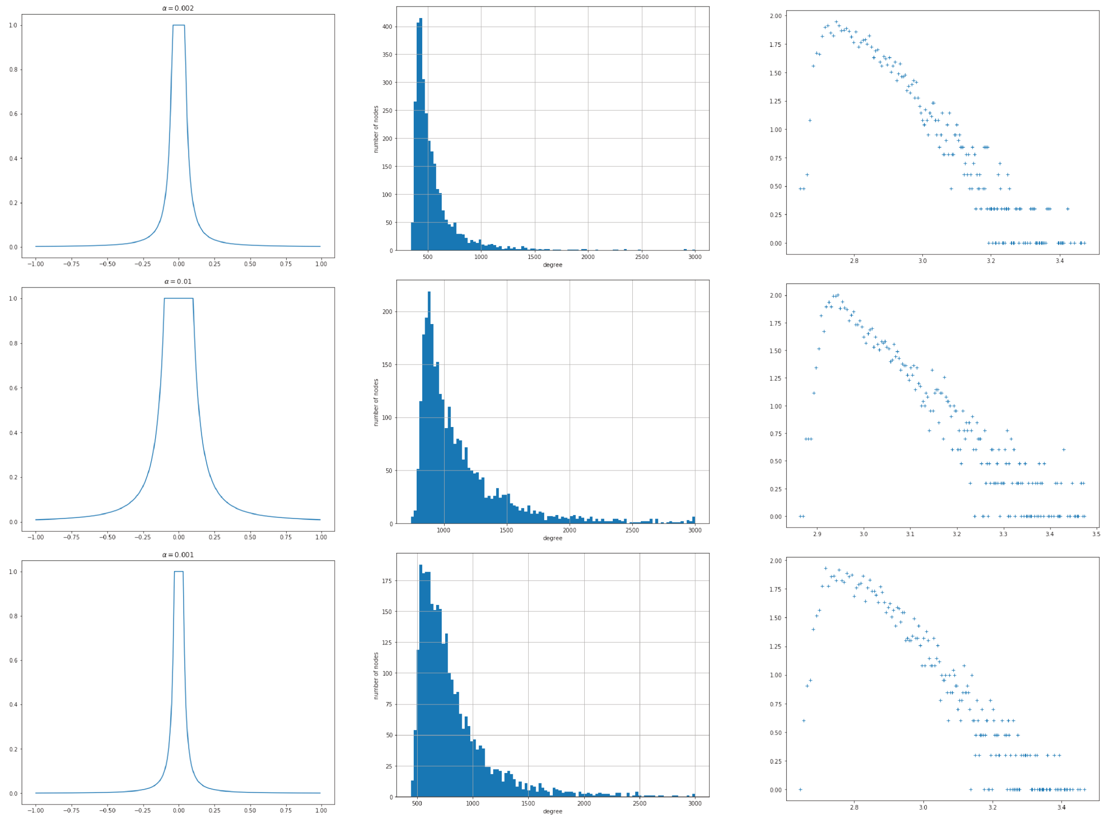

Figure 4.6: Analogous to Figure 4.5. In the log – log plot we show the values with degree larger than 300. The plot at the bottom was done with a different measure $F_{3/2}$

4.3 Latent norm recovery

As we saw in previous sections, the norm of a node X_i conveys important information about the connectivity (degree) of X_i under the RGG model on $\mathbb{B}^d$. We explore here the inverse of this relation, that is the inference of positional information (the norm) from the degree. Given that the points locations are typically not known and we only have access to the adjacency matrix, one can try to use the combinatorial information to construct an estimator for the norm. We present some results on this subject in this section.

First of all, we cannot recover the norm from the adjacency information in full generality, that is when considering the ensemble of all measures $\mathcal{F}$ and all the admissible link functions. To see that, we will consider two threshold graphons defining two models under two different measures in $\mathcal{F}$. If $\{X_i^\mu\}_{1\leq i\leq n}$ are i.i.d points on $\mathbb{B}^d$ with law μ and similarly $\{X_i^\nu\}$ are i.i.d points under ν, then the threshold graphons $W_1(\langle x,y\rangle)=\mathbb{1}_{\langle x,y\rangle>\tau_1}$ and $W_2(\langle x,y\rangle)=\mathbb{1}_{\langle x,y\rangle>\tau_2}$ will generate the same W-random graph models (the same probability distribution over finite graphs) if and only for any $i\neq j$

$$\mathbb{P}_\mu(\langle X_i^\mu, X_j^\mu\rangle \leq \tau_1)=\mathbb{P}_\nu(\langle X_i^\nu, X_j^\nu\rangle \leq \tau_2)$$

or, in words, if the τ_1 quantile of the inner products of points following μ is equal to the τ_2 quantile of the product under ν.

Proposition 44. *Let $\{X_i^\nu\}_{1\leq i\leq n}$ and $\{X_i^{\nu'}\}_{1\leq i\leq n}$ be two sets of points distributed under F_ν and $F_{\nu'}$ respectively for $\nu,\nu'>1/2$. Let τ be in $(0,1]$ and assume that $\nu>\nu'$, then we have*

$$\mathbb{P}_{\nu'}(\langle X_i^{\nu'}, X_j^{\nu'}\rangle \leq \tau)<\mathbb{P}_\nu(\langle X_i^\nu, X_j^\nu\rangle \leq \tau)$$

for $i\neq j$. Moreover, there exists $\tau'\in(0,1]$ such that

$$\mathbb{P}_{\nu'}(\langle X_i^{\nu'}, X_j^{\nu'}\rangle \leq \tau)=\mathbb{P}_\nu(\langle X_i^\nu, X_j^\nu\rangle \leq \tau')$$

for $i\neq j$.

Remark 9 (Case $\tau=0$)**.** *As we previously discussed, in the case $\tau=0$ we have equivalent models for any measure in $\mathcal{F}$. Indeed, for any measure with spherical symmetry we have the same model. Intuitively speaking, the norm is not used to decide the nodes connection, but only the fact that they belong to the same semisphere. In consequence, in the case $\tau=0$ we cannot recover the measure (nor distributional information about the latent points) from the adjacency matrix alone.*

The previous proposition proves that when no information about the measure is provided, we cannot extract information about the norm even if we restrict ourselves to the case of threshold graphons. Consequently, we cannot infer the latent norms from the combinatorial information alone. The previous can be seen as a non-linear analog (as in Prop. 44) of the non-identifiability issue under rescaling in the classic RGG on $\mathbb{R}^d$, which is reported in [Arias-Castro 2018] for instance. This is a common issue in latent space model and it is also present on the RDPG model.

Let fix $\nu>1/2$ and $\tau>0$. Recall that, for any node X_i, Lemma 43 relates the degree function with the norm $\|X_i\|$, given ν and τ. From that lemma, we get that $d_W(x)=0$ for any $\|x\|\leq\tau$. That is, the points inside $\mathbb{B}^d(0,\tau)$ are isolated (not connected with any other node) with probability 1. Notice that we can apply any function of $\mathbb{B}^d(0,\tau)$ to itself to the points in $\mathbb{B}^d(0,\tau)$ without changing the random graph distribution. In other words, we cannot extract interesting information on the points in $\mathbb{B}^d(0,\tau)$. Observe that it might be isolate nodes outside $\mathbb{B}^d(0,\tau)$, however the fraction of such points should go to zero asymptotically. We will return to this point later.

Fix a node such that $\|X_i\| > \tau$ and define the random variable

$$Z_i := I^{-1}(2\frac{d_G(X_i)}{n-1}; \nu + \frac{d}{2}, \frac{1}{2})$$

then Lemma 43 suggests the following estimator for the norm of X_i

$$\zeta_i := \frac{\tau}{\sqrt{1 - Z_i}}$$

From Lemma 43 we have that for a fixed $\alpha \in (0, 1)$ we have

$$|2\frac{d_G(X_i)}{n-1} - 2d_W(X_i)| \leq \sqrt{\frac{2\log 2/\alpha}{n}}$$

with probability larger than $1 - \alpha$. Otherwise stated, with probability larger than α we have

$$2\frac{d_G(X_i)}{n-1} \in [2d_W(X_i) - \sqrt{\frac{2\log 2/\alpha}{n}}, 2d_W(X_i) + \sqrt{\frac{2\log 2/\alpha}{n}}]$$

Given that $I^{-1}(\cdot)$ is strictly increasing, we have with probability larger than $1 - \alpha$

$$Z_i \in [I^{-1}(2d_W(X_i) - o(1)), I^{-1}(2d_W(X_i) + o(1))]$$

Observe that $2d_W(X_i)$ is strictly smaller than 1, because $\|X_i\| > \tau$. Thus, there exists $n_0 \in \mathbb{N}$ such that $2d_W(X_i) + o(1) < 1$ for $n \geq n_0$. This proves that the variable ζ_i is well defined for n large enough, with probability larger than $1 - \alpha$. To ensure that is well defined with probability 1 we take

$$\tilde{\zeta}_i = \zeta_i \wedge 1$$

From the strong law of large numbers it follows the following

Proposition 45. *For a fixed $i \in \mathbb{N}$, the random variable ζ_i converges almost surely to $\|X_i\|$.*

Even if the previous proves consistency of ζ_i, there is still some caveats to ζ_i as estimator. First, ζ_i is a complicated [3] non-linear function of a binomial random variable, which makes the analysis of its mean (and in consequence of the bias and variance) very involved. Second, the concentration properties are greatly distorted by the use of those nonlinear functions which deteriorates the rate of convergence. However, from our experiments (presented in Section 4.5) it do works reasonable well. Finding estimator with better properties, by possible considering not only local information about a node, is a line of research to be pursued in the future.

The convergence of the cumulative distribution of the degrees is proven in [Delmas 2018] and [Borgs 2018]. In [Borgs 2018], the authors prove the convergence of $|\{i \in [n] : d_G(X_i) > \lambda\}|$ towards $\mu(\{y : d_W(y) > \lambda\})$, where μ is the distributions of the points X_i and $\lambda > 0$ is a point of continuity of $\lambda \to \mu(\{d_W(y) > \lambda\})$. In [Delmas 2018], a graphon W in $[0, 1]$ is considered and the convergence of $\Pi(y) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{d_G(X_i) < nd_W(y)}$ toward y, almost surely, is proved. They also provide a CLT type result for this convergence.

[3] Since $I^{-1}(\cdot)$ cannot be expressed using standard functions.

4.4 Latent distances recovery

In this section we study the problem of latent distances estimation for the RGG on $\mathbb{B}^d$. The objective is to extend the spectral approach developed in Chapter 3 for the spherical case. One of the key ingredients used in our approach in the previous chapter was the explicit “reproducing" formula that relates the basis of Gegenbauer polynomial with the basis of eigenfunction of the integral operator with a dot product kernel. Analogous formulas for the case of the ball have been developed in the context of harmonic analysis on $\mathbb{B}^d$ [Dai 2013, Chap.11]. We will first introduce the harmonic analysis elements which are relevant to our method. In this section, we follow [Dai 2013, Chap.11] and [Xu 2001].

4.4.1 Harmonic analysis on the ball

Similar to the case of the sphere, we will use an orthogonal polynomials basis of $L^2(\mathbb{B}^d, F_\nu)$. Here the inner product, for $f, g \in L^2(\mathbb{B}^d, F_\nu)$ is given by

$$\langle f, g\rangle = a_\nu \int_{\mathbb{B}^d} f(x)g(x)dF_\nu(x)dx$$

where we recall that $dF_\nu(x) = (1-\|x\|^2)^{\nu-\frac{1}{2}}$ and $a_\nu = 1/\int_{\mathbb{B}^d} W_\nu(x)dx$.

We denote $\mathcal{Y}_n$ the subspace of orthogonal polynomials (with respect to the inner product defined above) of degree exactly n. It is implicit that $\mathcal{Y}_n$ depends on ν. From [Dai 2013, p.266], we know that the space dimension is $\dim \mathcal{Y}_n = \binom{n+d-1}{n}$ (this actually can be seen by applying a Gram-Schmidt orthonormalization process to monomials). This space can be described in terms of the harmonic polynomials of degree n on $\mathbb{S}^d$, identifying $x \in \mathbb{B}^d$ with $(x, \sqrt{1-\|x\|^2}) \in \mathbb{S}^d_+$, where $\mathbb{S}^d_+$ is the intersection of the sphere $\mathbb{S}^d$ with the half-space where the last coordinate is positive, see [Dai 2013, Lem.11.1.2]. Alternatively, they can be defined as the eigenfunctions of the differential operator

$$\mathcal{D} = \Delta - \langle x, \nabla\rangle^2 - (2\nu + d + 1)\langle x, \nabla\rangle$$

see [Dai 2013, Theorem 11.1.5], which is analogous to the case of the Laplace-Beltrami operator for the spherical harmonics on $\mathbb{S}^{d-1}$. However, we are not going to directly need these characterizations.

There are explicit expressions (closed formulas) for the reproducing kernel $P_n^\nu(x, y)$ of each $\mathcal{Y}_n$, see [Xu 2001]. In our context, the reproducing kernel $P_n^\nu(x, y)$ is the projector of $L^2(\mathbb{B}^d, F_\nu)$ onto $\mathcal{Y}_n$. By [Dai 2013, Cor.11.1.8] (see also [Xu 2001, Eq.2.2]) we have for $\nu > 0$, $\gamma_\nu := \nu + \frac{d-1}{2}$ and $c_\nu = \frac{\Gamma(\nu+1/2)}{\sqrt{\pi}\Gamma(\nu)}$

$$P_n^\nu(x, y) = c_\nu \frac{n+\gamma_\nu}{\gamma_\nu} \int_{-1}^{1} G_n^{\gamma_\nu}(\langle x, y\rangle + \sqrt{1-\|x\|^2}\sqrt{1-\|y\|^2}t)(1-t^2)^{\nu-1}dt \quad (4.2)$$

where $G_n^{\gamma_\nu}(\cdot)$ is the Gegenbauer polynomial of degree n with weight γ_ν.

It is well known, and we used it in the previous chapter, that $\{G_n^{\gamma_\nu}(\cdot)\}_{n\geq 0}$ forms a basis for $L^2([-1,1],\gamma_\nu)$ [Szego 1939]. In addition, each $p_k \in \mathcal{Y}_n$ is an eigenfunction of the following $L^2(\mathbb{B}^d, W_\nu)$ integral operator

$$T_f g(x) = \int_{\mathbb{B}^d} f(\langle x, y\rangle) g(y) W_\nu(y) dy$$

and the eigenvalue associated to each $p_k \in \mathcal{Y}_n$ is

$$\lambda_n^*(f) = c_{\gamma_\nu} \int_{-1}^{1} f(t) \frac{G_n^{\gamma_\nu}(t)}{G_n^{\gamma_\nu}(1)} (1-t^2)^{\gamma_\nu - 1/2} dt$$

and c_{γ_ν} is such that $\lambda_0^*(1) = 1$. The previous statement is a consequence of the Funk-Hecke formula [Dai 2013, Thm.11.1.9]. Notice that for each $n \in \mathbb{N}$ we have the following decomposition of the reproducing kernel of $\mathcal{Y}_n$ in terms of the basis elements p_k

$$P_n^\nu(x,y) = \sum_{p_k \in \mathcal{Y}_n} p_k(x) p_k(y) \tag{4.3}$$

and for a given $f(\langle x, y\rangle)$ the following decomposition holds

$$f(\langle x, y\rangle) = \sum_{n\in\mathbb{N}} \lambda_n^*(f) P_n^\nu(x,y) \tag{4.4}$$

Formula (4.3) tell us how to reconstruct the reproducing kernel of each eigenspace of T_f from the elements of the basis. On the other hand, formula (4.2) give us the explicit form for each reproducing kernel. For the linear polynomial $G_1^\gamma(t) = 2\gamma t$, we have that

$$\begin{aligned} P_1^\nu(x,y) &= 2c_\nu \frac{1+\gamma_\nu}{\gamma_\nu} \int_{-1}^{1} (\langle x, y\rangle + \sqrt{1-\|x\|^2}\sqrt{1-\|y\|^2}t)(1-t^2)^{\nu-1} dt \\ &= 2\tilde{c_\nu}(1+\gamma_\nu)\langle x, y\rangle \end{aligned} \tag{4.5}$$

where $\tilde{c_\nu} = c_\nu \int_{-1}^{1}(1-t^2)^{\nu-1}dt$. In the last step we used that $\int_{-1}^{1} t(1-t^2)^{\nu-1}dt = 0$ given the parity of the function inside the integral. From formula (4.3) we deduce that

$$\frac{1}{2\tilde{c_\nu}(1+\gamma_\nu)} \sum_{p_k \in \mathcal{Y}_n} p_k(x) p_k(y) = \langle x, y\rangle \tag{4.6}$$

The previous relation is the analogous to the Eq.3.4 in the case of the dot product kernels on $\mathbb{S}^{d-1}$. In order to have similar results we need to study the finite sample properties of the spectra of the graphon $W(x,y) = f(\langle x, y\rangle)$ and its associated probability matrix and adjacency matrix.

4.4.2 Graphon eigensystem and estimation result

We recall the notation from Chapters 2 and 3. We consider the integral operator $T_W : L^2(\mathbb{B}^d) \to L^2(\mathbb{B}^d)$

$$T_W g(x) = \int_{\mathbb{B}^d} f(\langle x, y\rangle) g(y) dF_\nu(y)$$

and the $n \times n$ symmetric matrices

$$T_n = \frac{1}{n}(1-\delta_{ij})f(\langle X_i, X_j\rangle)$$
$$\hat{T}_n = \frac{1}{n}A_{ij}$$

where A_{ij} is the adjacency matrix defined in Section 4.2. A few known results about the spectral behavior for this objects

- Koltchinskii-Giné result [Koltchinskii 2000, Thm.1] is formulated in an abstract space and it holds in this case, given that the kernel is square integrable. Then we have a.s

  $$\delta_2(\lambda(T_n), \lambda(T_W)) \to 0$$

- We can invoke Bandeira-Van Handel result [Bandeira 2016, Cor.3.12] to prove that $\lambda(\hat{T}_n)$ is close to $\lambda(T_n)$ for n large. We recall that using Thm. 32 we obtain

  $$\|\hat{T}_n - T_n\|_{op} \lesssim_\alpha \frac{1}{\sqrt{n}}$$

  with probability larger than $1-\alpha$.

- In the spherical case developed in the previous chapter we use the result [De Castro 2020, Thm.2]. In [De Castro 2020] there is a more a general result which holds in this case. Alternatively, we can use Cor. 5 in Chapter 2 which we recall gives that, with probability larger than $1-\alpha$

  $$\delta_2(\lambda(T_n), \lambda(T_W)) = \mathcal{O}_\alpha(\frac{1}{\sqrt{n}})$$

  provided that the regularity parameter δ (using the notation used throughout Chapters 2 and 3) is larger than $2s+1$. In this case we need to obtain an estimate of the parameter s as we did for the spherical case in Section 2.7.

Recalling that L^2 basis of eigenfunctions of T_W is given by $\cup_{n\geq 0}\cup_{p_k\in\mathcal{Y}_n} p_k$. We have following estimates for the sup-norm.

Lemma 46. *We have for $p_k \in \mathcal{Y}_n$*

$$\|p_k\|_\infty \lesssim n^{\nu+\frac{d-1}{2}} \qquad \textit{for } 1 \leq k \leq \dim(\mathcal{Y}_n)$$
$$\Big\| \sum_{k=1}^{\dim(\mathcal{Y}_n)} p_k^2 \Big\|_\infty \lesssim n^{2\nu+d-1}$$

Moreover, we have

$$\mathcal{V}_1(i) = \mathcal{O}_\nu(i^{\frac{2\nu+d}{d}})$$

where $\mathcal{V}_1(\cdot)$ is the variance proxy parameter defined in Section 2.3.

Recalling the definition of Sobolev regularity in Section 2.6, we will say that $f : [-1,1] \to \mathbb{R}$ is γ_ν-Sobolev regular of parameter s, or $f \in S^p_{\gamma_\nu}([-1,1])$, if f satisfies that

$$\|f\|_{p,\gamma_\nu} := \sqrt{\sum_{l\geq 0} \lambda_l^*(f)^2 (1 + l^p(l+2\nu+d-1)^p)} \leq \infty$$

More on Weighted Sobolev spaces can be found in [Nicaise 2000]. It is direct from the definition that if $f \in S^p_{\gamma_\nu}([-1,1])$ then $\lambda_l^*(f) = O(l^{-p-1-\varepsilon})$ for $\varepsilon > 0$. The values $l^p(l+2\nu+d-1)$ appearing in the definition of the Sobolev space are the eigenvalues of the differential operator $\mathcal{D}$.

We recall and update some of the notation of the previous chapter. We call $\mathcal{G}^* = \frac{1}{n}(1-\delta_{ij})\langle X_i, X_j\rangle$ the population Gram matrix and $\mathcal{G}_U := \frac{1}{2\tilde{c}_\nu(1+\gamma_\nu)} UU^T$ for any $n \times d$ real matrix U. The reason for dividing by the constant $2\tilde{c}_\nu(1+\gamma_\nu)$, comes from Eq. (4.6).

Theorem 47. *Let $p > 2\nu - 1 + \frac{5d}{2}$ and W be a graphon defined by a dot product kernel on $\mathbb{B}^d$ and measure F_ν. Assume that $W \in S^p_{\gamma_n u}([-1,1])$ satisfy the spectral gap condition $\Delta^* > 0$, then there exists a set of d eigenvalues $\hat{v}_1, \cdots, \hat{v}_d$ of $\hat{T}_n$ such that we have with probability larger than $1-\alpha$*

$$\|\mathcal{G}^* - \hat{\mathcal{G}}\|_F = \mathcal{O}_\alpha(\Delta^{*-1}\frac{1}{\sqrt{n}})$$

where $\hat{\mathcal{G}} = \mathcal{G}_{\hat{V}}$ and $\hat{V}$ is the matrix with columns $\hat{v}_1, \hat{v}_2, \cdots, \hat{v}_d$.

Remark 10. *The regularity condition $p > 2\nu - 1 + \frac{5d}{2}$ is mainly technical. We imposed it in order to use Cor. 5 and obtain parametric rate. Under the weaker condition $p > \nu + \frac{d-1}{2}$ we have uniform convergence of the kernel spectral expansion and we can use [De Castro 2020, Thm.2] plugging the estimates from Lemma 46 and optimizing the truncation parameter R we obtain a rate for $\delta_2(\cdot,\cdot)$ given by*

$$\delta_2(\lambda(T_n), \lambda(T_W)) \leq n^{\frac{-\delta+1}{2\delta-1+\nu(1+\frac{1}{d})+\frac{d}{2}}}$$

where $\delta = \frac{1}{2} + \frac{p}{d-2}$. In the proof of Thm.47, as in the proof of Thm.26, the main ingredient to determine the rate for the error $\|\mathcal{G}^* - \hat{\mathcal{G}}\|_F$ is the rate for $\delta_2(\cdot,\cdot)$. Observe that when $\delta \to \infty$, and the rest of parameters are fixed, the right hand side converges to $n^{-\frac{1}{2}}$.

4.5 Numerical Experiments

We run simulation for different RGG models on $\mathbb{B}^d$ and compute the different estimators analyzed throughout this chapter to see how they perform empirically. In the case of the latent distances estimation, we run the algorithm HEiC, which corresponds to Algorithm 1 described in Section 3.6. The only change is that the constant for which we multiply the sum of the outer product of the d eigenvalues in the cluster of λ_1^*, now is $\frac{1}{2\tilde{c}_\nu(1+\gamma_\nu)}$ instead of $\frac{1}{d}$.

4.5.1 Latent norm recovery

We study the empirical performance of the estimator ζ, for which we prove almost sure converge to the latent norm on the threshold RGG model. We compute the estimator for each node in the graph, following measure of error for each sample

$$E_{norm} = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{\|X_i\| \geq \tau}} \Big(\sum_{i=1}^{n} (\zeta_i - \|X_i\|)^2 \mathbb{1}_{\|X_i\| \geq \tau} \Big)^{1/2}$$

We discard the points with norm larger than τ, because as we discussed in Section 4.3 the adjacency matrix of the graph carries no information about the norm of those points, other than being smaller than τ. In Figure 4.7(left) we plot the mean square error E_{norm} in logarithmic scale for a threshold $\tau = 0.1$. For each sample size, we simulate 25 graphs on the ball with dimension $d = 3$, and uniform measure $F_{1/2}$, and compute the mean of the errors. The form in which the error decrease suggest a parametric rate of convergence, which we plot in a red line for reference. However, note that the fact the estimator is based in a complicated nonlinear function, as it is

$$t \to \frac{\tau}{\sqrt{1 - I^{-1}(t; \frac{d}{2} + \nu, \frac{1}{2})}}$$

makes that this rate is non-uniform across the nodes. Indeed, given the shape of the graph of $I^{-1}(t; \frac{d}{2} + \nu, \frac{1}{2})$ it is not hard to see that points in with higher norm (closer to 1) will converge slower. This indeed what we observe in the experiments as shown in Figure 4.7(right), where we plot the sequence of ordered norms in red and the sequence of ordered $\zeta_i's$ for different values of the sample size ($n = 100,$). Notice that it takes much more samples to see a convergence when the norm of the node is close to 1.

We observe that for values of τ closer to 0, the convergence is indeed slower. In Figure 4.7 we plot the mean of $\log(E_{norm})$ over 25 sampled graphs, for a threshold $\tau = 0.01$ with dimension parameter $d = 3$ and the measure $F_{1/2}$. We observe that it takes much more samples to converge. Even if the decay of the errors suggest a similar parametric rate in the case of the model with smaller τ, the constant (intercept) is larger, which means that the error is always larger than in the previous case. This should not be surprising given that we know that in the model with $\tau = 0$ we cannot infer the norm from the samples (as the mode, is equivalent to the threshold graphon on the sphere). Approaching to $\tau = 0$ will render the problem harder, in the sample complexity sense.

To see empirically the effect of changing the parameter ν in the estimation of the norm, in Figure 4.9 we plot the mean of the error $\log(E_{norm})$ across 25 samples for the threshold graphon with $\tau = 0.1$ and $d = 3$. We see that a larger ν gives lower error, this is explained by the fact the larger the ν, the more concentrated the sampled nodes are close to the center of the ball. This can be seen by the distribution of the norm (squared) which is plotted in Figure 4.9 (right).

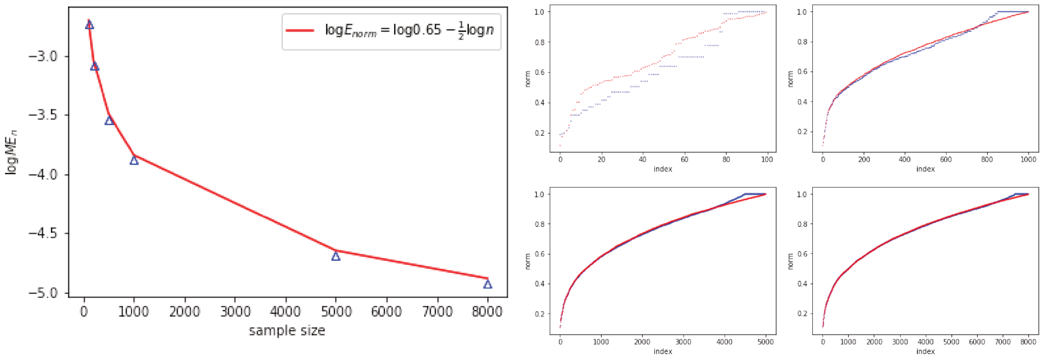

Figure 4.7: In the left we show the mean of $\log(E_{norm})$ over 25 graphs, for the recovery of the norm on the threshold graph with $\tau = 0.1$ and parameter $d = 3$ and measure $F_{1/2}$. We add the upper in red for reference. In the right we plot the sequence of ordered values for $\|X_i\|$ in increasing order together with the sorted sequence of estimated values ζ_i.

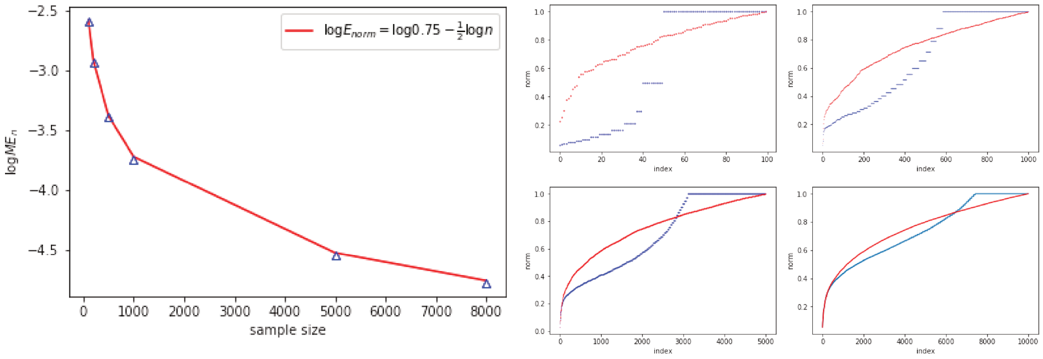

Figure 4.8: This is analogous to Figure 4.7, with $\tau = 0.01$ and maintaining the rest of parameters.

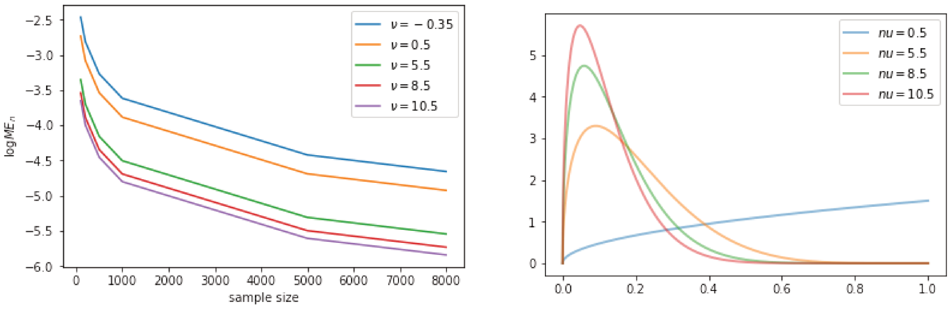

Figure 4.9: Similar to Figure 4.8. We plot the error for the threshold graphon with $\tau = 0.1$, $d = 3$ and changing the measure F_ν. In the right we plot the pdf of $Beta(\frac{d}{2}, \nu + \frac{1}{2})$ which is the distribution of the norm of the nodes.

4.5.2 Latent distance recovery

We report the empirical performance of the algorithm HEiC, described in Section 3.6 applied in the context of RGG in $\mathbb{B}^d$. Similar to the spherical case, we will mainly measure the mean error

$$ME_n = \|\mathcal{G}^* - \mathcal{G}\|_F$$

We first consider the threshold graphon with parameter $\tau = 0.1$ in dimension $d = 3$. We sample 25 graphs using this model and run each time the algorithm HEiC. In Figure 4.10(left) we show a boxplot for $\log(ME_n)$ for different sample sizes. In Figure 4.10(right) we show the $\log(ME_n)$ error for different values of n in the case of the logistic graphon $f(t) = \frac{1}{1+e^{rt}}$ for different values of r. The curves in the plot were obtaining by averaging across 25 samples for each value of n. We observe that for $r = -0.1$ the error does not decrease with the sample size, which is to be expected as the logistic function for that value of r is close to a constant function. In our parametrization of the problem, this translate into a close to 0 spectral gap as the Figure 4.11 illustrates. Indeed, we plot the first 10 eigenvalues, for this case the cluster of eigenvalues associated to λ_1^* is a subset of the first 10 eigenvalues. We see that as r is closer to zero, the spectral gap decrease, and the number of samples required to have a decreasing error, increase.

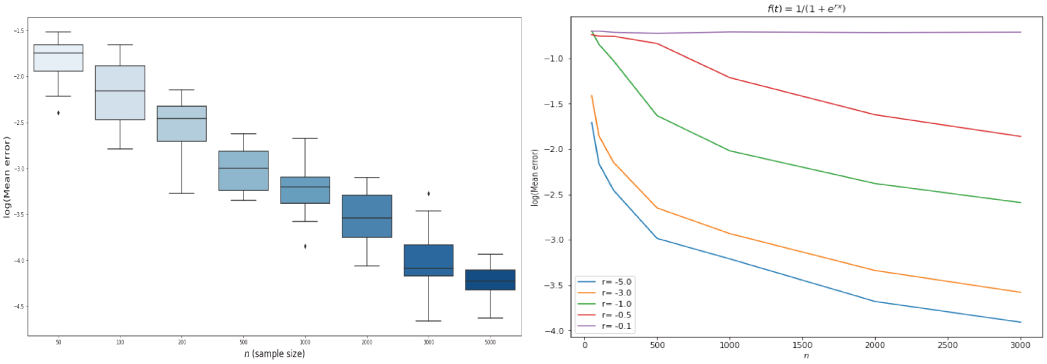

Figure 4.10: In the left a boxplot for $\log(ME_n)$, for different values of the sample size in threhsold graphon with $d = 3$, $\tau = 0.1$ and $F_{1/2}$. In the right we plot the mean error for the logistic graphon with different values of the parameter r.

Note that Theorem 47 do not give information about the diagonal of the Gram matrix, which corresponds to the square of the norms of the nodes X_i. Our measure of error ME_n do not take them into consideration. In the case of the threshold graphon we can use the estimator ζ_i to compute the means. We observe empirically that the algorithm works better when the rows matrix of eigenvectors $\hat{V}$, which has columns $v_1, \cdots, v_d$ which are the output of the algorithm HEiC, are normalized to match the mean of the true means ζ_i. This is not an ideal situation from the practical point of view, given that the norms are usually non available. In the case of the threshold graphon we can use the estimated norms in this extra normalization

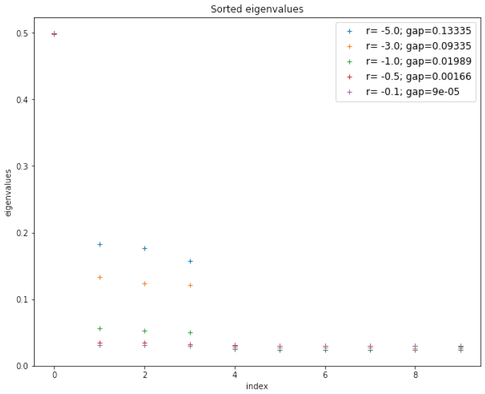

Figure 4.11: We plot the first 10 eigenvalues for the logistic graphon for different values of the parameter r. We include the spectral gap in each case. In all the examples we used a parameter $n = 1000$.

step. While this gives reasonable results in practice, a thorough theoretical study is lacking at this moment and will be left for future work.

Remark 11. *The time complexity(or running time) of the latent distance recovery algorithm does not increase, in comparison with the spherical case, and it is roughly $\mathcal{O}(n^3+n)$. In the case of the computation of the estimators ζ_i we need to compute the degrees, which corresponds to compute the sum of all rows, which is roughly $\mathcal{O}(n^2)$.*

4.6 Comments and future work

Throughout this chapter we have studied the problem of estimating latent structures for graphs generated by the RGG model on $\mathbb{B}^d$. In practice, some of the estimators we have presented here, such as ζ_i, suffers problems related to its theoretical guarantees and its rate of convergence. The fact that they are defined as a complicated function of the data makes harder a direct finite sample analysis. We expect that the use of global information, in conjunction with the degree function, would help in finding simpler estimators which are more prone to be studied with the standard concentration tools.

The problem of estimating τ, given the information of μ, is also of interest. This problem has been studied in the model with $\Omega = [-1, 1]^2$ in [Diaz 2018], with the uniform distribution. They propose an estimator based on the count of common neighbors for a small set of subsampled nodes (forced to be not so close to have conditional independency). In the model we presented here, we believe that simpler estimators are possible, given the fact isolated nodes give information about $F_\nu(\mathbb{B}^d(0, \tau))$. The main difficulty will be to estimate, with high probability, the number of isolated nodes whose associated points are outside $\mathbb{B}^d(0, \tau)$. Constructing such estimators in left for future work.

4.7 Useful results

We state and prof some of the most relevant results for this section, many of which deals with geometric formulas

Lemma 48 (Threshold graphon degree density)**.** *Let W be the threshold graphon and f the probability density function of $d_W(X)$, where $X \sim F_\nu$ we have for $t > 0$*

$$f(t) = \frac{\tau^d}{(1 - I^{-1}(2t))^{\frac{d}{2}+1}} (1 - \frac{\tau^2}{1 - I^{-1}})^{\nu - \frac{1}{2}} \frac{1}{I(2t)^{\frac{d}{2}+\nu}(1 - I(2t))^{-\frac{1}{2}}} \tag{4.7}$$

where we use the notation $I(t) = I_t(\frac{d}{2} + \nu, \frac{1}{2})$.

Proof. It is well known that the function $t \to I_t(a, b)$ is differentiable and it is straightforward to check that $g(t)$ it is also differentiable for $t > 0$. Taking the derivative of $F_{d_W}(t) = I_{g(t)}(\frac{d}{2}, \nu + \frac{1}{2})$ the result follows from simple computations

$$\begin{aligned} f_g(t) &= \frac{1}{B(\frac{d}{2}, \nu + \frac{1}{2})} g(t)^{\frac{d}{2}-1}(1 - g(t))^{\nu - \frac{1}{2}} g'(t) \\ &= \frac{\tau^d}{(1 - I^{-1}(2t))^{\frac{d}{2}+1}} (1 - \frac{\tau^2}{1 - I^{-1}})^{\nu - \frac{1}{2}} \frac{1}{I(2t)^{\frac{d}{2}+\nu}(1 - I(2t))^{-\frac{1}{2}}} \end{aligned}$$

□

The following result gives a characterization for a basis of $\mathcal{V}_l$. The proof can be found in [Dai 2013, Thm.11.1.12]

Theorem 49. *The space $\mathcal{V}_n$ has a basis consisting on functions $G_l^{\gamma_\nu}(\langle x, \psi_i \rangle)$ for some points $\{\psi_i\}_{1 \le i \le \dim(\mathcal{V}_l)}\} \subset \mathbb{S}^{d-1}$.*

4.8 Some proofs

Here we gather some proofs or proof sketches for the main results of this section.

Proof of Lemma 40 . It is classic (see [Kelker 1970, Sec.5]) that for a spherically symmetric distribution with density of the form $p(y) = g(\|y\|^2)$ where $y \in \mathbb{B}^d$, then the norm will have density $h(r) = \frac{2\pi^{d/2}}{\Gamma(d/2)} r^{d-1} g(r^2)$. The c.d.f for the radius of variable distributed following F_ν is proportional to $\int_0^t r^{d-1}(1 - r^2)^{\nu - \frac{1}{2}} dr$ using the change of variables $r \to r^2$ we obtain that the square of the norm have density $\int_0^t r^{\frac{d-1}{2}}(1 - r)^{\nu - \frac{1}{2}} dr$ where we recognize the density of a $Beta(\frac{d}{2}, \nu + \frac{1}{2})$. □

Proof of Prop.42. Notice that in the case of the threshold graphon, the degree function $t \to d_W(tN)$ is increasing. Using this we have that

$$\mathbb{P}(d_W(t_1 N)) \le d_W(X_i) \le d_W(t_2 N) = \mathbb{P}(\|X_i\| \in [t_1, t_2])$$

Using the previous and Lemma 40, the result follows. □

Proof of Lemma 43. For $t \leq \tau$ we have that $tN \in \mathbb{B}^d(0,\tau)$, which implies that $d_W(tN) = 0$. The result for this case follows by noting that $I_0(a,b) = 0$ for any $0 \leq a, b \leq 1$. For $t > \tau$, call $h = (\tau/t)^2$ we have

$$\begin{aligned} d_W(tN) &= \int_{\mathbb{B}^d} \mathbb{1}_{\langle tN,y\rangle \geq \tau}(1-\|y\|^2)^{\nu-\frac{1}{2}} \\ &\propto \int_h^1 \int_0^{\sqrt{1-x_1^2}} r^{d-2}\big(1-x_1^2-r^2\big)^{\nu-\frac{1}{2}} dr dx_1 \\ &\propto \int_h^1 (1-x_1)^{\frac{d}{2}+\nu-1} dx_1 \int_0^1 (1-t)^{\nu-\frac{1}{2}} t^{\frac{d-3}{2}} dt \\ &\propto \int_0^{1-h} x_1^{\frac{d}{2}+\nu-1} x_1^{\frac{1}{2}} dx_1 \end{aligned}$$

where we did a change a change of variables $t = \frac{r^2}{1-x_1^2}$ in the third line. The result follows from the fact the both quantities integrate 1. □

Proof sketch Prop.45. Fix $i \in [n]$, from the strong law of large numbers we have that $\frac{1}{n-1} d_G(X_i)$, which conditionally to X_i is a sum of independent random variables, converges to $d_W(X_i)$. The continuity of $I^{-1}(\cdot; a, b)$ implies that $I^{-1}(\frac{1}{n-1} d_G(X_i); a, b)$ converges to $I^{-1}(d_W(X_i), a, b)$. The follows by using Lemma 43 and properties of $I(\cdot; a, b)$.

□

Proof of Lemma 46 . The first inequality comes from the basis characterization in Thm. 49. Not considering normalization constants we have for p_k in $\mathcal{V}_l$

$$\begin{aligned} \|p_k\|_\infty &\propto \|G_l^{\gamma_\nu}(\langle \cdot, \psi_i \rangle)\|_\infty \\ &\lesssim \|G_l^{\gamma_\nu}\|_\infty \end{aligned}$$

From [Szego 1939, Thm.7.32.1] we deduce that $\|G_l^{\gamma_\nu}\|_\infty = G_l^{\gamma_\nu}(1)$ and from [, Cor.3.2] we obtain the estimate $G_l^{\gamma_\nu}(1) \asymp n^{\gamma_\nu - \frac{1}{2}}$. Replacing $\gamma_\nu = \frac{d-1}{2} + \nu$ the result follows.

For the second inequality we use the addition theorem $\sum_{p_k \in \mathcal{V}_l} p_k(x) p_k(y) = P_l(x,y)$ and Eq.(4.2). We have

$$\begin{aligned} \sum_{p_k \in \mathcal{V}_l} p_k^2(x) = P_l(x,x) \\ \propto \int_{-1}^1 G_l^{\gamma_\nu}(\|x\|^2 + (1-\|x\|^2)t)(1-t^2)^{\nu-1} dt \end{aligned}$$

The result follows by using the same estimates used to prove the first inequality. □

Proof sketch Thm. 47. The ideas for the proof of Thm. 47 go in the same line that Thm. 26, but using the results in Sec. 3.9, instead of the estimates in [De Castro 2020]. The condition $p > 2\nu + d + 1$ comes from the estimates in Lemma

46 and the fact that we use Cor. 5. This gives a bound for the rate of convergence of the eigenvalues. Noting that in the proof of Thm. 26 we only use the specific form of the graphon in the eigenvalues estimates and in the recovery formula, which in case has it analog in Eq.(4.2). □

Bibliography

[Abbe 2017] E. Abbe, F. Baccelli et A. Sankararaman. *Community Detection on Euclidean Random Graphs.* arXiv:1706.09942, 2017. (Cité en pages 9 et 48.)

[Amini 2020] A. Amini et Z. Razaee. *Concentration of kernel matrices with application to kernel spectral clustering.* arXiv:1909.03347, 2020. (Cité en page 18.)

[Araya 2019] E. Araya et Y. De Castro. *Latent distance estimation for random geometric graph.* Advances in Neural Information Processing Systems, pages 8721–8731, 2019. (Cité en pages 2, 32, 47 et 53.)

[Araya 2020] E. Araya. *Relative concentration bound for the spectrum of kernel matrices.* ArXiv, 2020. (Cité en pages 2, 8, 14, 18, 19, 21, 26, 31 et 43.)

[Arbel 2017] J. Arbel et O. Marchal. *On the sub-Gaussianity of the Beta and Dirichlet distributions.* Electron. Commun. Probab., vol. 22, no. 54, 2017. (Cité en page 43.)

[Arbel 2019] J. Arbel, O. Marchal et H. Nguyen. *On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables.* arXiv:1901.09188, 2019. (Cité en page 43.)

[Arcones 1993] M Arcones et E. Giné. *Limit theorems for U-processes.* Annals of Probability., vol. 21, no. 3, pages 1494–1542, 1993. (Cité en page 40.)

[Arias-Castro 2014] E. Arias-Castro et N. Verzelen. *Community detection in dense random networks.* Annals of Statistics, vol. 42, no. 3, pages 940–969, 2014. (Cité en pages 9 et 73.)

[Arias-Castro 2015] E. Arias-Castro et N. Verzelen. *Community detection in sparse random networks.* Annals of Applied Probability, vol. 25, no. 6, pages 3465–3510, 2015. (Cité en pages 9 et 73.)

[Arias-Castro 2018] E. Arias-Castro, A. Channarond, B. Pelletier et N. Verzelen. *On the estimation of latent distances using graph distances.* arXiv:1804.10611, 2018. (Cité en pages 48, 54 et 80.)

[Athreya 2018] A. Athreya, D.E. Fishkind, M. Tang, C. Priebe, Y. Park, J. Vogelstein, K. Levin, V. Lyzinski, Y. Qin et D. Sussman. *Statistical Inference on Random Dot Product Graphs: a Survey.* Journal of Machine Learning Research, vol. 18, pages 1–92, 2018. (Cité en page 69.)

[Athreya 2020] A. Athreya, M. Tang, Y. Park et C. Priebe. *On estimation and inference in latent structure random graphs.* arXiv:1806.01401, 2020. (Cité en page 70.)

[Bandeira 2016] A. Bandeira et R. Van Handel. *Sharp nonasymptotic bounds on the norm of random matrices with independent entries.* Annals of Probability, vol. 44, no. 4, pages 2479–2506, 2016. (Cité en pages 17, 51, 61, 67 et 84.)

[Barrat 2004] A. Barrat, R. Barthélemey, R. Pastor-Satorras et A. Vespignani. *The architecture of complex weighted networks.* PNAS, vol. 101, no. 11, pages 3747–3752, 2004. (Cité en pages 70 et 75.)

[Bathia 1997] R. Bathia. Matrix analysis. Springer Verlag New York, 1997. (Cité en pages 18, 61, 65 et 66.)

[Bednorz 2014] W. Bednorz. *Concentration via chaining method and its applications.* arXiv:1405.0676, 2014. (Cité en page 42.)

[Belkin 2018] M. Belkin. *Approximation beats concentration?* COLT.Proceedings of Machine Learning Research, vol. 75, pages 1–14, 2018. (Cité en pages 6, 14, 18, 21, 26 et 27.)

[Blanchard 2007] G. Blanchard, O. Bousquets et L. and Zwald. *Statistical properties of Kernel PCA.* Machine Learning, vol. 66, pages 259–294, 2007. (Cité en pages 7, 13, 18, 19 et 26.)

[Borgs 2008] C. Borgs, J.T Chayes, L. Lovasz, V.T. Sos et K. Vesztergombi. *Convergent sequences of dense graphs I: subgraph frequencies, metric properties, and testing.* Adv. Math, vol. 219, pages 1802–1852, 2008. (Cité en pages 3 et 14.)

[Borgs 2010] C. Borgs, J.T Chayes et L. Lovasz. *Moments of two-variable functions and the unique-ness of graph limits.* Geom.Funct. Anal, vol. 19, pages 1597–1619, 2010. (Cité en pages 3 et 14.)

[Borgs 2012] C. Borgs, J.T Chayes, L. Lovasz, V.T. Sos et K. Vesztergombi. *Convergent sequences of dense graphs II. Multiway cuts and statistical physics.* Annals of Mathematics, vol. 176, no. 1, pages 151–219, 2012. (Cité en pages 3, 14 et 31.)

[Borgs 2018] C. Borgs, J. Chayes, H. Cohn et N. Holden. *Sparse Exchangeable Graphs and Their Limits via Graphon Processes.* Journal of Machine Learning Research, vol. 18, pages 1–71, 2018. (Cité en page 81.)

[Borgs 2019] C. Borgs, J. Chayes, H. Cohn et Y. Zhao. *An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions.* Trans. Amer. Math. Soc., vol. 372, pages 3019–3062, 2019. (Cité en page 77.)

[Boucheron 2013] S. Boucheron, G. Lugosi et P. Massart. Concentration inequalities: A non asymptotic theory of independence. Oxford University Press, first édition, 2013. (Cité en pages 19 et 39.)

[Braun 2005] M. Braun. *Spectral Properties of the Kernel Matrix and their Relation to Kernel Methods in Machine Learning.* PhD thesis, Rheinische Friedrich-Wilhelms-Universität, Bonn, 2005. (Cité en page 39.)

[Braun 2006] M. Braun. *Accurate error bounds for the eigenvalues of the kernel matrix.* Journal of Machine Learning Research, vol. 7, pages 2303–2328, 2006. (Cité en pages 6, 7, 8, 18, 21, 22, 26 et 27.)

[Bubeck 2016] S. Bubeck, J. Ding, R. Eldan et M. Rácz. *Testing for high dimensional geometry in random graphs.* Random Structures and Algorithms, vol. 49, pages 503–532, 2016. (Cité en pages 9, 32, 34, 48 et 73.)

[Bubeck 2017] S. Bubeck et M. Rácz. *Basic models and questions in statistical network analysis.* Statistics surveys, vol. 11, pages 1–47, 2017. (Cité en pages 32 et 36.)

[Cao 2019] Y Cao, Z. Fang, Y. Wu, D. Zhou et Q. Gu. *Towards Understanding the Spectral Bias of Deep Learning.* arXiv:1912.01198, 2019. (Cité en page 14.)

[Chatterjee 2015] S. Chatterjee. *Matrix estimation by Universal Singular Value Thresholding.* Annals of Statistics, vol. 43, no. 1, pages 177–214, 2015. (Cité en pages 49, 51, 53 et 56.)

[Clauset 2009] A. Clauset, C.R Shalizi et M.E.J Newman. *Power law distributions in empirical data.* SIAM review, vol. 51, no. 4, pages 661–703, 2009. (Cité en pages 11, 75, 77 et 78.)

[Cunningham 2017] W. Cunningham, K. Zuev et D. Krioukov. *Navigability of Random Geometric Graphs in the Universe and Other Spacetime.* Scientific reports, vol. 7, page art.num. 8699, 2017. (Cité en pages 32 et 48.)

[Dai 2013] F. Dai et Y. Xu. Approximation theory and harmonic analysis on spheres and balls. Springer Verlag Monographs in Mathematics, 2013. (Cité en pages 29, 30, 52, 70, 82, 83 et 90.)

[De Castro 2020] Y. De Castro, C. Lacour et T.M Pham Ngoc. *Adaptive estimation of nonparametric geometric graphs.* Mathematical Statistics and Learning, 2020. (Cité en pages 6, 8, 9, 18, 21, 22, 31, 32, 51, 54, 62, 66, 84, 85 et 91.)

[De La Pena 1995] V.H De La Pena et S.J. Montgomery-Smith. *Decoupling inequalities for the tail probabilities of multivariate U-statistics.* Annals of Probability, vol. 23, no. 2, pages 806–816, 1995. (Cité en page 40.)

[Delmas 2018] J.F. Delmas, J.S. Dhersin et M. Sciauveau. *Asymptotic for the cumulative distribution function of the degrees and homomorphism densities for random graphs sampled from a graphon.* arXiv:1807.09989, 2018. (Cité en page 81.)

[Diaconis 1981] P. Diaconis et D. Freedman. *The statistics of vision: the Julesz conjecture.* J.Math.Psychology, vol. 2, pages 112–138, 1981. (Cité en page 3.)

[Diaz 2018] J. Diaz, C. McDiarmid et D. Mitsche. *Learning random points from geometric graphs or orderings.* arXiv:1804.10611, 2018. (Cité en pages 48, 54 et 89.)

[Dirksen 2015] S. Dirksen. *Tail bounds via generic chaining.* Electronic Journal of Probability, vol. 20, no. 53, 2015. (Cité en page 42.)

[El Karoui 2010] N. El Karoui. *The spectrum of kernel random matrices.* Ann. Probab., vol. 38, no. 1, pages 1–50, 2010. (Cité en page 17.)

[Emery 1998] M. Emery, A. Nemirovski et D. Voiculescu. Lectures on probability theory,. Springer-Verlag Berlin Heidelberg, Ecole d'ete de probabilites de saint-flour XXVIII édition, 1998. (Cité en page 54.)

[Eren 2017] T. Eren. *The effects of random geometric graph structure and clustering on localizability of sensor networks.* International Journal of Distributed Sensor Networks, vol. 13, no. 12, 2017. (Cité en pages 9 et 48.)

[Fasshauer 2011] G.E Fasshauer. *Positive definite kernels: past, present and future.* Dolomite Research Notes on Approximation, no. 4, pages 21–63, 2011. (Cité en page 38.)

[Franceschetti 2008] M Franceschetti et R Meester. Random networks for communication: from statistical physics to information systems. Cambridge University Press, first édition, 2008. (Cité en page 31.)

[Gilbert 1961] E.N. Gilbert. *Random plane networks.* J.Soc.Industrial Applied Mathematics, vol. 9, no. 5, pages 533–543, 1961. (Cité en pages 4, 48 et 49.)

[Gine 2000] E. Gine, R. Latala et J. Zinn. *Exponential and Moment Inequalities for U-Statistics.* High Dimensional Probability II, pages 13–38, 2000. (Cité en pages 8, 24 et 40.)

[Gine 2015] E. Gine et R. Nickl. Mathematical foundation of infinite-dimensional statistical models. Cambridge University Press, 2015. (Cité en page 40.)

[Higham 2008a] D. Higham, M. Rasajski et N. Przulj. *Fitting a geometric graph to a protein-protein interaction network.* Bioinformatics, vol. 24, no. 8, page 1093?1099, 2008. (Cité en page 48.)

[Higham 2008b] D.J. Higham, M. Rasajski et N. Przulj. *Fitting a geometric graph to a protein-protein interaction network.* Bioinformatics, vol. 24, no. 8, pages 1093–1099, 2008. (Cité en page 32.)

[Hirsch 1999] F. Hirsch et G. Lacombe. Elements of functional analysis. Springer-Verlag New York, 1999. (Cité en pages 14 et 15.)

[Hoff 2002] P. Hoff, A. Raftery et M. Handcock. *Latent space approaches to social network analysis.* Journal of the American Statistical Association, vol. 97, no. 460, pages 1090–1098, 2002. (Cité en pages 37 et 48.)

[Hofmann 2008] T. Hofmann, B. Schölkopf et Smola A.J. *Kernel methods in machine learning.* Annals of statistics, vol. 36, no. 3, pages 1171–1220, 2008. (Cité en page 13.)

[Horn 2012] R. Horn et C. Johnson. Matrix analysis. Cambridge University Press, 2012. (Cité en page 39.)

[Houdré 2003] C. Houdré et P. Reynaud-Bouret. *Exponential inequalities, with constants, for U-statistics of order 2.* Stochastic inequalities and applications, pages 55–69, 2003. (Cité en pages 24 et 40.)

[Indritz 2019] J. Indritz. *An inequality for Hermite polynomials.* Proceedings of the American Mathematical Society, vol. 12, no. 6, page 981?983, 2019. (Cité en page 38.)

[Ipsen 1998] I. Ipsen. *Relative perturbation results for matrix eigenvalues and singular values.* Acta Numerica, vol. 7, pages 151–201, 1998. (Cité en pages 7 et 18.)

[Javanmard 2013] A. Javanmard et A. Montanari. *Localization from Incomplete Noisy Distance Measurements.* Foundations of computational mathematics, vol. 13, page 297?345, 2013. (Cité en page 48.)

[Jia 2004] X. Jia. *Wireless networks and random geometric graphs.* Proc. Int. Symp. Parallel Architectures, Algorithms and Networks, pages 575–579, 2004. (Cité en page 48.)

[Jirak 2019] M. Jirak et M. Wahl. *Perturbation bounds for eigenspaces under a relative gap condition.* Proc. Amer. Math. Soc., vol. 448, 2019. (Cité en page 18.)

[Kasiviswanathan 2015] S.P. Kasiviswanathan et M. Rudelson. *Spectral Norm of Random Kernel Matrices with Applications to Privacy.* arXiv preprint arXiv:1504.05880, 2015. (Cité en pages 14 et 18.)

[Kato 1995] T. Kato. *Perturbation Theory for Linear Operators.* Classics in Mathematics (132). Springer, 2nd ed. 1995. (Cité en page 16.)

[Kelker 1970] D. Kelker. *Theory of spherical distributions and a location-scale parameter generalization.* The indian journal of statistics, Serie A., vol. 32, no. 4, pages 419–430, 1970. (Cité en pages 71 et 90.)

[Klopp 2017a] O. Klopp, A. Tsybakov et N. Verzelen. *Oracle inequalities for network models and sparse graphon estimation.* Annals of Statistics, vol. 45, no. 1, pages 316–354, 2017. (Cité en pages 14, 32 et 49.)

[Klopp 2017b] O. Klopp et N. Verzelen. Annals of Statistics, vol. 45, no. 1, pages 316–354, 2017. (Cité en page 4.)

[Koltchinskii 1998] V. Koltchinskii. *Asymptotics of spectral projections of some random matrices approximating integral operators.* Progress in Probability, vol. 43, no. In: Eberlein E., Hahn M., Talagrand M. (eds) High Dimensional Probability, pages 191–227, 1998. (Cité en pages 50 et 51.)

[Koltchinskii 2000] V. Koltchinskii et E. Giné. *Random matrix approximation of spectra of integral operators.* Bernoulli, pages 113–167, 2000. (Cité en pages 5, 8, 14, 17, 19, 21, 22, 25, 50 et 84.)

[Koltchinskii 2017] V. Koltchinskii et Loucini. *Concentration inequalities and moment bounds for sample covariance operators.* Bernoulli, vol. 23, no. 1, pages 110–133, 2017. (Cité en pages 13, 18, 26, 40, 41 et 42.)

[Ledoux 1991] M. Ledoux et M. Talagrand. Probability in banach spaces: isoperimetry and processes. Berlin:Springer, 1991. (Cité en page 42.)

[Levin 2017] K. Levin et V. Lyzinski. *Laplacian eigenmaps from sparse, noisy similarity measurements.* IEEE Transactions on Signal Processing, vol. 65, pages 1998–2003, 2017. (Cité en pages 10 et 53.)

[Li 2009] J Li, L. Andrew, C. Heng Foh, M. Zuckerman et H. Chen. *Connectivity, Coverage and Placement in Wireless Sensor Networks.* Sensor (Basel), vol. 9, no. 10, page 7664?7693, 2009. (Cité en pages 9 et 48.)

[Lounici 2019] K. Lounici. *High-dimensional covariance matrix estimation with missing observations.* Bernoulli, vol. 448, 2019. (Cité en pages 18, 26 et 27.)

[Lovász 2006a] L. Lovász et B. Szegedy. *Limit of dense graph sequences.* J.Combin. Theory Ser.B, vol. 98, pages 933–957, 2006. (Cité en pages 3 et 14.)

[Lovász 2006b] L. Lovász et B. Szegedy. *Limits of dense graph sequences.* J.Combin.Theory.Ser B, vol. 96, no. 6, pages 197–215, 2006. (Cité en pages 3 et 14.)

[Lovasz 2012] L. Lovasz. Large networks and graph limits. Colloquium Publications (AMS), 2012. (Cité en pages 3, 4, 32, 37, 49, 50 et 71.)

[Marchenko 1967] V.A Marchenko et L.A Pastur. *Distribution of eigenvalues for some sets of random matrices.* Mat. Sb. N.S. (in Russian), vol. 72, no. 114:4, pages 507–536, 1967. (Cité en page 17.)

[Mckey 2014] L. Mckey, M. Jordan, R.Y. Chen, B. Farrell et J. Tropp. *Matrix concentration inequalities via the method of exchangeable pairs.* Ann. Probab., vol. 42, no. 3, pages 906–945, 2014. (Cité en page 23.)

[Mckey 2016] L. Mckey et J. Tropp. *Efron-Stein inequalities for random matrices.* Ann.Probab., vol. 44, no. 5, pages 3431–3473, 2016. (Cité en page 23.)

[McPherson 2001] M. McPherson, L. Smith-Lovin et J.M. Cook. *Birds of a Feather: Homophily in Social Networks.* Annual Review of Sociology, vol. 27, pages 415–444, 2001. (Cité en page 48.)

[Mitzenmacher 2003] M. Mitzenmacher. *A Brief History of Generative Models for Power Law and Lognormal Distributions.* Internet Math., vol. 1, no. 2, pages 226–251, 2003. (Cité en pages 11 et 75.)

[Nicaise 2000] S. Nicaise. *Jacobi polynomials, weighted Sobolev spaces and approximation results of some singularities.* Math. Nachr., vol. 213, pages 117–140, 2000. (Cité en pages 28 et 85.)

[Oh 2010] S. Oh, A. Montanari et Amin Karbasi. *Sensor Network Localization from Local Connectivity : Performance Analysis for the MDS-MAP Algorithm.* IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo), 2010. (Cité en page 48.)

[Ostovskii 2019] D. Ostovskii et A. Rudi. *Affine Invariant Covariance Estimation for Heavy-Tailed Distributions.* COLT, 2019. (Cité en pages 18, 26 et 27.)

[Penrose 2003] M. Penrose. Random geometric graphs. Oxford University Press, first édition, 2003. (Cité en pages 32 et 36.)

[Penrose 2016] M. Penrose. *Connectivity of soft random geometric graphs.* Annals of Applied Probability, vol. 26, no. 2, pages 986–1028, 2016. (Cité en page 9.)

[Rácz 2019] M. Rácz et J. Richey. *A Smooth Transition from Wishart to GOE.* Journal of Theoretical Probability, vol. 32, pages 898–906, 2019. (Cité en page 9.)

[Rasmussen 2006] C.E Rasmussen et C. Williams. Gaussian processes for machine learning. The MIT Press., 2006. (Cité en page 33.)

[Rosasco 2010] L. Rosasco, M. Belkin et E. De Vito. *On learning with integral operators.* Journal of Machine Learning Research, vol. 11, pages 905–934, 2010. (Cité en pages 6, 7, 16, 21 et 26.)

[Sarkar 2010] P. Sarkar, D. Chakrabarti et A.W. Moore. *Theoretical justification of popular link prediction heuristics.* International Conference on Learning Theory, 2010. (Cité en pages 9 et 48.)

[Shawe-Taylor 2005] J. Shawe-Taylor, C. Williams, N. Cristiani et J. Kandola. *On the Eigenspectrum of the Gram Matrix and the Generalization Error of Kernel PCA.* IEEE Transactions on Information Theory, vol. 51, no. 7, pages 2510–2522, 2005. (Cité en pages 18, 19 et 26.)

[Sussman 2014] D.L. Sussman, M. Tang et C.E. Priebe. *Consistent latent position estimation and vertex classification for random dot product graphs.* IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 36, pages 48–57, 2014. (Cité en page 9.)

[Szego 1939] G. Szego. Orthogonal polynomials. Colloquium Publications (AMS), 1939. (Cité en pages 33, 38, 83 et 91.)

[Talagrand 1996] M. Talagrand. *Majorizing measures: the generic chaining.* Annals of Probability, vol. 24, no. 3, pages 1049–1103, 1996. (Cité en page 42.)

[Tang 2013] M. Tang, D.L Sussman et C.E Priebe. *Universally consistent vertex classification for latent position graphs.* Annals of Statistics, vol. 41, pages 1406–1430, 2013. (Cité en pages 48 et 53.)

[Tropp 2012] J Tropp. *User-friendly tail bounds for sums of random matrices.* Foundations of Computational Mathematics, vol. 12, no. 4, pages 389–434, 2012. (Cité en pages 8 et 39.)

[Van Handel 2017] R. Van Handel. *Structured random matrices.* Convexity and Concentration (Carlen et al., eds.), IMA., vol. 161, pages 107–165, 2017. (Cité en page 24.)

[Vershynin 2012a] R. Vershynin. *How close is the sample covariance matrix to the actual covariance matrix?* Journal of Theoretical Probability, vol. 25, pages 655–686, 2012. (Cité en pages 8, 18, 40 et 62.)

[Vershynin 2012b] R. Vershynin. *Introduction to the non-asymptotic analysis of random martices.* In: Compressed sensing, Theory and Applications. Edited by Y.Eldar and G. Kurtyniok, Chap. 5, pages 210–268, 2012. (Cité en page 40.)

[Vershynin 2018] R. Vershynin. High-dimensional probability: An introduction with applications in data science. Cambridge University Press, 2018. (Cité en page 62.)

[Xu 2001] Y. Xu. *Representation of Reproducing Kernels and the Lebesgue Constants on the Ball.* Journal of Approximation Theory, vol. 112, pages 295–310, 2001. (Cité en page 82.)

[Xu 2017] J Xu. *Rate of Convergence of Spectral Methods for Graphon Estimation.* arXiv:1709.03183, 2017. (Cité en page 28.)

[Yu 2015] Y. Yu, T. Wang et R.J. Samworth. *A useful variant of the Davis-Kahan theorem for statisticians.* Biometrika, vol. 102, no. 2, pages 315–323, 2015. (Cité en page 60.)

[Zhou 2002] D.X Zhou. *The Covering Number in Learning Theory.* Journal of Complexity, vol. 18, pages 739–767, 2002. (Cité en page 41.)

[Zhu 1998] H. Zhu, C. Williams, R. Rohwer et M. Morcinie. *Gaussian regression and optimal finite dimensional linear models.* in Neural networks and machine learning, C. Bishop, ed., 1998. (Cité en page 38.)

université
PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Apprentissage spectral des noyaux et inférence des graphes aléatoires géométriques

Mots clés: matrices à noyau, graphon, graphes aléatoires géométriques, espace latent

Résumé: Cette thèse comporte deux objectifs. Un premier objectif concerne l'étude des propriétés de concentration des matrices à noyau, qui sont fondamentales dans l'ensemble des méthodes à noyau. Le deuxième objectif repose quant à lui sur l'étude des problèmes d'inférence statistique dans le modèle des graphes aléatoires géométriques. Ces deux objectifs sont liés entre eux par le formalisme du graphon, qui permet représenter un graphe par un noyau. Nous rappelons les rudiments du modèle du graphon dans le premier chapitre. Le chapitre 2 présente des bornes précises pour les valeurs propres individuelles d'une matrice à noyau, où notre principale contribution est d'obtenir des inégalités à l'échelle de la valeur propre en considération. Ceci donne des vitesses de convergence qui sont meilleures que la vitesse paramétrique et, en occasions, exponentielles. Jusqu'ici cela n'avait été établi que avec des hypothèses contraignantes dans le contexte des graphes. Nous spécialisons les résultats au cas de noyaux de produit scalaire, en soulignant sa relation avec le modèle des graphes géométriques. Le chapitre 3 étudie le problème d'estimation des distances latentes pour le modèle des graphes aléatoires géométriques dans la sphère Euclidienne. Nous proposons un algorithme spectral efficace qui utilise la matrice d'adjacence pour construire un estimateur de la matrice des distances latentes, et des garanties théoriques pour l'erreur d'estimation, ainsi que la vitesse de convergence, sont montrées. Le chapitre 4 étend les méthodes développées dans le chapitre précédent au cas des graphes aléatoires géométriques dans la boule Euclidienne, un modèle qui, en dépit des similarités formelles avec le cas sphérique, est plus flexible en termes de modélisation. En particulier, nous montrons que pour certaines choix des paramètres le profile des dégrées est distribué selon une loi de puissance, ce qui a été vérifié empiriquement dans plusieurs réseaux réels. Tout les résultats théoriques des deux dernier chapitres sont confirmés par des expériences numériques.

Title: Kernel spectral learning and inference in random geometric graphs

Keywords: kernel matrices, graphon model, random geometric graph, latent space

Abstract: This thesis has two main objectives. The first is to investigate the concentration properties of random kernel matrices, which are central in the study of kernel methods. The second objective is to study statistical inference problems on random geometric graphs. Both objectives are connected by the graphon formalism, which allows to represent a graph by a kernel function. We briefly recall the basics of the graphon model in the first chapter. In chapter two, we present a set of accurate concentration inequalities for individual eigenvalues of the kernel matrix, where our main contribution is to obtain inequalities that scale with the eigenvalue in consideration, implying convergence rates that are faster than parametric and often exponential, which hitherto has only been establish under assumptions which are too restrictive for graph applications. We specialized our results to the case of dot products kernels, highlighting its relation with the random geometric graph model. In chapter three, we study the problem of latent distances estimation on random geometric graphs on the Euclidean sphere. We propose an efficient spectral algorithm that use the adjacency matrix to construct an estimator for the latent distances, and prove finite sample guaranties for the estimation error, establishing its convergence rate. In chapter four, we extend the method developed in the previous chapter to the case of random geometric graphs on the Euclidean ball, a model that despite its formal similarities with the spherical case it is more flexible for modelling purposes. In particular, we prove that for certain parameter choices its degree profile is power law distributed, which has been observed in many real life networks. All the theoretical findings of the last two chapters are verified and complemented by numerical experiments.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France