ASYMPTOTIQUES POUR LA FONCTION DE RÉPARTITION EMPIRIQUE DES DEGRÉS ET LES DENSITÉS D'HOMOMORPHISMES DE GRAPHES ALÉATOIRES ÉCHANTILLONNÉS À PARTIR D'UN GRAPHON

Version légèrement modifiée de l'article [53]
Asymptotics for the cumulative distribution of the degrees and homomorphism densities for random graphs sampled from a graphon
soumis pour publication.

Abstract. We give asymptotics for the cumulative distribution function (CDF) for degrees of large dense random graphs sampled from a graphon. The proof is based on precise asymptotics for binomial random variables.

Replacing the indicator function in the empirical CDF by a smoother function, we get general asymptotic results for functionals of homomorphism densities for partially labeled graphs with smoother functions. This general setting allows to recover recent results on asymptotics for homomorphism densities of sampled graphon.

4.1 Introduction

The Internet, social networks or biological networks can be represented by large random graphs. Understanding their structure is an important issue in Mathematics. The degree sequence is one of the key objects used to get informations about graphs. The degree sequences of real world networks have attracted a lot of attention during the last years because their distributions are significantly different from the degree distributions studied in the classical models of random graphs such as the Erdős-Rényi model where the degree distribution is approximately Poisson when the number of nodes is large. They followed a power-law distribution, see for instance, Newmann [158], Chung et al [49], Diaconis and Blitzstein [26] and Newman, Barabasi and Watts [155]. See also Molloy and Reed [147, 148] and Newman, Strogatz and Watts [156] in the framework of sparse graphs.

In this paper, we shall consider the cumulative distribution function (CDF) of degrees of large dense random graphs sampled from a graphon, extending results from Bickel, Chen and Levina [25]. The theory of graphon or limits of sequence of dense graphs was developed by Lovász and Szegedy [139] and Borg, Chayes, Lovász, Sós and Vesztergombi [40]. The asymptotics on the empirical CDF of degrees, see the theorem in Section 4.1.1, could be used to test if a large dense graph is sampled from a given graphon. This result is a first step for giving a non-parametric test for identifying the degree function of a large random graph in the spirit of the Kolmogorov-Smirnov test for the equality of probability distribution from a sample of independent identically distributed random variables.

If we replace the indicator function in the empirical CDF by a smoother function, we get general results on the fluctuations for functionals of homomorphism densities for partially labeled graphs. As an application, when considering homomorphism densities for sampled graphon, we recover results from Féray, Méliot and Nikeghbali [84].

4.1.1 Convergence of CDF of empirical degrees for large random graphs

We consider simple finite graphs, that is graphs without self-loops and multiple edges between any pair of vertices. We denote by $\mathcal{F}$ the set of all simple finite graphs.

There exists several equivalent notions of convergence for sequences of finite dense graphs (that is graphs where the number of edges is close to the maximal number of edges), for instance in terms of metric convergence (with the cut distance) or in terms of the convergence of subgraph densities, see [40] or Lovász [138].

When it exists, the limit of a sequence of dense graphs can be represented by a graphon i.e. a symmetric, measurable function $W : [0,1]^2 \to [0,1]$, up to a measure preserving bijection. A graphon W may be thought of as the weight matrix of an infinite graph whose set of vertices is the continuous unit interval, so that $W(x,y)$ represents the weight of the edge between vertices x and y.

Moreover, it is possible to sample simple graphs, with a given number of vertices, from a graphon W (called W-random graphs). Let $X = (X_i : i \in \mathbb{N}^*)$ be a sequence of independent random variables uniformly distributed on the interval $[0,1]$. To construct the W-random graph with vertices $[n] := \{1, \ldots, n\}$, denoted by G_n, for each pair of distinct vertices $i \neq j$, elements of $[n]$, connect i and j with probability $W(X_i, X_j)$, independently of all other edges (see also Section 4.2.4). If needed, we shall stress the dependence on W and write $G_n(W)$ for G_n. By this construction, we get a sequence of random graphs $(G_n : n \in \mathbb{N}^*)$ which converges almost surely towards the graphon W, see for instance Proposition 11.32 in [138].

We define the degree function $D = (D(x) : x \in [0,1])$ of the graphon W by:

$$D(x) = \int_0^1 W(x,y)dy.$$

And we consider the empirical CDF $\Pi_n = (\Pi_n(y) : y \in [0,1])$ of the normalized degrees of the graph G_n defined by

$$\Pi_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\left\{D_i^{(n)} \leq D(y)\right\}},$$

where $(n-1)D_i^{(n)}$ is the degree of the vertex i in G_n.

Bickel, Chen and Levina [25], Theorem 5 (with $m = 1$), proved the convergence in distribution and the convergence of the second moments of $\Pi_n(y)$ towards y. We improve the results given in [25]: under the condition that D is increasing[1] on $[0,1]$, we have the almost sure convergence of $\Pi_n(y)$ towards y, uniformly on $[0,1]$. This is a consequence of the more general result given by Theorem 4.9 (see Subsection 4.1.2 and Remark 4.22 for more details).

[1]Since the graphon is defined up to a measure preserving one-to-one map on $[0,1]$, there exists an equivalent version of the graphon for which the degree function is non-decreasing. If the degree function is increasing, then this version is unique in L^1 and this is the version which is considered in this section.

In a different direction, Chatterjee and Diaconis [45] considered the convergence of uniformly chosen random graphs with a given CDF of degrees towards an exponential graphon with given degree function.

We also get the fluctuations associated to the almost sure convergence of Π_n. If W satisfies some regularity conditions given by (4.62), which in particular imply that D is of class $\mathcal{C}^1$, then we have the following result on the convergence in distribution of finite-dimensional marginals for Π_n.

Theorem (Theorem 4.23)**.** *Assume that W satisfies condition (4.62). Then we have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\,(\Pi_n(y) - y) : y \in (0,1)\right) \xrightarrow[n\to+\infty]{(fdd)} \chi,$$

where $(\chi_y : y \in (0,1))$ is a centered Gaussian process defined, for all $y \in (0,1)$ by:

$$\chi_y = \int_0^1 (\rho(y,u) - \bar{\rho}(y))dB_u,$$

with $B = (B_u, u \geq 0)$ a standard Brownian motion, and $(\rho(y,u) : u \in [0,1])$ and $\bar{\rho}(y)$ defined for $y \in (0,1)$ by:

$$\rho(y,u) = \mathbf{1}_{[0,y]}(u) - \frac{W(y,u)}{D'(y)} \quad \text{and} \quad \bar{\rho}(y) = \int_0^1 \rho(y,u)du.$$

The covariance kernel $\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_3$ of the Gaussian process χ is explicitly given by Equations (4.64), (4.65) and (4.66) which define respectively Σ_1, Σ_2 and Σ_3. In particular, we deduce that the variance of $\chi(y)$, for $y \in (0,1)$ is given by the elementary formula:

$$\Sigma(y,y) = y(1-y) + \frac{1}{D'(y)^2}\left(\int_0^1 W(y,x)^2dx - D(y)^2\right) + \frac{2}{D'(y)}\left(D(y)y - \int_0^y W(y,x)dx\right).$$

The proof of this result relies on uniform Edgeworth expansions for binomial random variables, see Bhattacharya and Rao [22], and Stein's method for binomial random vectors, see Bentkus [20]. The convergence of the process in the Skorokhod space could presumably be proved using similar but more involved arguments. More generally, following van der Vaart [188], Chapter 19 on convergence of empirical CDF of independent identically distributed random variables, if would be natural to study the uniform convergence of $\frac{1}{n}\sum_{i=1}^n f(D_i^{(n)})$ when f belongs to a certain class of functions.

The asymptotics on the CDF of empirical degrees appear formally as a limiting case of the asymptotics of $\frac{1}{n}\sum_{i=1}^n f(D_i^{(n)})$ with f smooth. This is developed in Section 4.1.3. We shall in fact adopt in this section a more general point of view as we replace the normalized degree sequence by a sequence of homomorphism densities for partially labeled graphs.

4.1.2 Convergence of sequence of dense graphs towards graphons

Recall that one of the equivalent notions of convergence of sequences of dense graphs is given by the convergence of subgraph densities. It is the latter one that will interest us. We first recall the notion of homomorphism densities. For two simple finite graphs F and G with respectively $v(F)$ and $v(G)$ vertices, let $\mathrm{Inj}(F,G)$ denote the set of injective homomorphisms (injective adjacency-preserving maps) from F to G (see Subsection 4.2.2 for a precise definition). We define the injective homomorphism density from F to G by the following normalized quantity:

$$t_{\mathrm{inj}}(F,G) = \frac{|\mathrm{Inj}(F,G)|}{A_{v(G)}^{v(F)}},$$

where we have for all $n \geq k \geq 1$, $A_n^k = n!/(n-k)!$. In the same way, we can define the density of induced homomorphisms (which are injective homomorphisms that also preserve non-adjacency), see (4.24). Some authors study subgraph counts rather than homomorphism densities, but the two quantities are related, see Bollobás and Riordan [32], Section 2.1, so that results on homomorphism densities can be translated into results for subgraph counts.

A sequence of dense simple finite graphs $(H_n : n \in \mathbb{N}^*)$ is called convergent if the sequence $(t_{\mathrm{inj}}(F,H_n) : n \in \mathbb{N}^*)$ has a limit for every $F \in \mathcal{F}$. The limit can be represented by a graphon, say W and we have that for every $F \in \mathcal{F}$:

$$\lim_{n\to\infty} t_{\mathrm{inj}}(F,H_n) = t(F,W),$$

where

$$t(F,W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i,x_j) \prod_{k\in V(F)} dx_k.$$

According to [138], Proposition 11.32, the sequence of W-random graphs $(G_n : n \in \mathbb{N}^*)$ converges a.s. towards W, that is for all $F \in \mathcal{F}$, a.s.:

$$\lim_{n\to\infty} t_{\mathrm{inj}}(F,G_n) = t(F,W). \tag{4.1}$$

In the Erdős-Rényi case, that is when $W \equiv \mathfrak{p}$ is constant, the fluctuations associated to this almost sure convergence are of order n: for all $F \in \mathcal{F}$ with p vertices and e edges, we have the following convergence in distribution:

$$n\left(t_{\mathrm{inj}}(F,G_n(\mathfrak{p})) - \mathfrak{p}^e\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, 2e^2\mathfrak{p}^{2e-1}(1-\mathfrak{p})\right),$$

where $\mathcal{N}(m,\sigma^2)$ denotes a Gaussian random variable with mean m and variance σ^2. There are several proofs of this central limit theorem. Nowicki [159] and Janson and Nowicki [118] used the theory of U-statistics to prove the asymptotic normality of subgraph counts and induced subgraph counts. They also obtained the asymptotic normality of vectors of subgraph counts and induced subgraph counts. In the particular case of the joint distribution of the count of edges, triangles and two-stars, Reinert and Röllin [167], Proposition 2, obtained bounds on the approximation. Using discrete Malliavin calculus, Krokowski and Thäle [132] generalized the result of [167] (in a different probability metric) and get the rate of convergence associated to the multivariate central limit theorem given in [118]. See also Féray, Méliot and Nikeghbali [83], Section 10, for the mod-Gaussian convergence of homomorphism densities.

The asymptotics of normalized subgraph counts have also been studied when the parameter $\mathfrak{p}$ of the Erdős-Rényi graphs depends on n, see for example Ruciński [174], Nowicki and Wierman [161], Barbour, Karoński and Ruciński [18], and Gilmer and Kopparty [100].

In the general framework of graphons, the speed of convergence in the invariance principle is of order $\sqrt{n}$, except for degenerate cases such as the Erdős-Rényi case. This result was given by Féray, Méliot and Nikeghbali [84], Theorem 21: for all $F \in \mathcal{F}$, we have the following convergence in distribution:

$$\sqrt{n}\left(t_{\mathrm{inj}}(F,G_n) - t(F,W)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma(F)^2\right), \tag{4.2}$$

where, with $V(F)$ the set of vertices of F and $v(F)$ its cardinality,

$$\sigma(F)^2 = \sum_{q,q' \in V(F)} t\big((F \bowtie F)(q,q'), W\big) - v(F)^2 \, t(F,W)^2$$

and $(F \bowtie F')(q,q')$ is the disjoint union of the two simple finite graphs F and F' where we identify the vertices $q \in F$ and $q' \in F'$ (see point (iii) of Remark 4.12, for more details). Notice that in the Erdős-Rényi case, that is when W is a constant graphon, the asymptotic variance $\sigma(F)^2$ is equal to 0, which is consistent with the previous paragraph since the speed is of order n.

Using Stein's method, Fang and Röllin [81] obtained the rate of convergence for the multivariate normal approximation of the joint distribution of the normalized edge count and the corrected and normalized 4-cycle count. As a consequence, they get a confidence interval to test if a given graph G comes from an Erdős-Rényi random graph model or a non constant graphon-random graph model. Maugis, Priebe, Olhede and Wolfe [143] gave a central limit theorem for subgraph counts observed in a network sample of W-random graphs drawn from the same graphon W when the number of observations in the sample increases but the number of vertices in each graph observation remains finite. They also get a central limit theorem in the case where all the graph observations may be generated from different graphons. This allows to test if the graph observations come from a specified model. When considering sequences of graphons which tend to 0, then there is a Poisson approximation of subgraph counts. In this direction, Coulson, Gaunt and Reinert [51], Corollary 4.1, used the Stein method to establish an effective Poisson approximation for the distribution of the number of subgraphs in the graphon model which are isomorphic to some fixed strictly balanced graph.

Motivated by those results, we present in the next section an invariance principle for the distribution of homomorphism densities of partially labeled graphs for W-random graphs which can be seen as a generalization of (4.2).

4.1.3 Asymptotics for homomorphism densities of partially labeled graphs for large random graphs

Let $n \in \mathbb{N}^*$ and $k \in [n]$. We define the set $\mathcal{S}_{n,k}$ of all $[n]$-words of length k such that all characters are distinct, see (4.7). Notice that $|\mathcal{S}_{n,k}| = A_n^k = n!/(n-k)!$.

We generalize homomorphism densities for partially labeled graphs. Let $F, G \in \mathcal{F}$ be two simple graphs with $V(F) = [p]$ and $V(G) = [n]$. Assume $n \geq p > k \geq 1$. Let $\ell \in \mathcal{S}_{p,k}$ and $\alpha \in \mathcal{S}_{n,k}$. We define $\mathrm{Inj}(F^\ell, G^\alpha)$ the set of injective homomorphisms f from F into G such that $\mathrm{f}(\ell_i) = \alpha_i$ for all $i \in [k]$, and its density:

$$t_{\mathrm{inj}}(F^\ell, G^\alpha) = \frac{|\mathrm{Inj}(F^\ell, G^\alpha)|}{A_{n-k}^{p-k}}.$$

We define the random probability measure $\Gamma_n^{F,\ell}$ on $([0,1], \mathcal{B}([0,1]))$, with $\mathcal{B}([0,1])$ the Borel σ-field on $[0,1]$, by: for all measurable non-negative function g defined on $[0,1]$,

$$\Gamma_n^{F,\ell}(g) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} g\left(t_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right). \tag{4.3}$$

We prove, see Theorem 4.9, the almost sure convergence for the weak topology of the sequence $\left(\Gamma_n^{F,\ell}(dx) : n \in \mathbb{N}^*\right)$ of random probability measure on $[0,1]$ towards a deterministic probability measure $\Gamma^{F,\ell}(dx)$, see definition (4.42).

- If we take $g = \mathrm{Id}$ in (4.3), we recover the almost sure convergence given in (4.1) as according to (4.21):
$$t_{\mathrm{inj}}(F, G_n) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\mathrm{inj}}(F^\ell, G_n^\alpha).$$

- If we take $g = \mathbf{1}_{[0,D(y)]}$ with $y \in (0,1)$ and $F = K_2$ (where K_2 denotes the complete graph with two vertices) in (4.3) and using the expression of $\Gamma^{F,\ell}$ given in Remark 4.8, (ii), we have, with $\bullet$ any vertex of K_2, that:
$$\Gamma_n^{K_2,\bullet}(g) = \Pi_n(y) \quad \text{and} \quad \Gamma^{K_2,\bullet}(g) = y.$$
Then, by Theorem 4.9, under the condition that D is strictly increasing on $(0,1)$, we have the almost sure convergence of $\Pi_n(y)$ towards y, see Remark 4.22.

We also have the fluctuations associated to this almost sure convergence, see Theorem 4.11 for a multidimensional version.

Theorem. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}$ be a simple finite graphs with $V(F) = [p]$, $\ell \in \mathcal{M}_p$, with $k = |\ell|$. Then, for all $g \in \mathcal{C}^2([0,1])$, we have the following convergence in distribution:*
$$\sqrt{n}\left(\Gamma_n^{F,\ell}(g) - \Gamma^{F,\ell}(g)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^{F,\ell}(g)^2\right),$$
with $\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(\mathcal{U}_g^{F,\ell}\right)$ and $\mathcal{U}_g^{F,\ell}$ is defined in (4.43).

Notice $\sigma^{F,\ell}(g)^2$ is an integral involving g and g'. The asymptotic results are still true when we consider a family of $d \geq 1$ simple graphs $F = (F_m : 1 \leq m \leq d) \in \mathcal{F}^d$ and we define $\Gamma_n^{F,\ell}$ on $[0,1]^d$, see Theorems 4.9 and 4.11 for the multidimensional case. The case $g = \mathrm{Id}$ appears already in [84], see Corollary 4.13 for the graphs indexed version. We have the following convergence of finite-dimensional distributions (or equivalently of the process since $\mathcal{F}$ is countable).

Corollary (Corollary 4.13)**.** *We have the following convergence of finite-dimensional distributions:*
$$\left(\sqrt{n}\,(t_{\mathrm{inj}}(F, G_n) - t(F, W)) : F \in \mathcal{F}\right) \xrightarrow[n\to\infty]{(fdd)} \Theta_{\mathrm{inj}},$$
where $\Theta_{\mathrm{inj}} = (\Theta_{\mathrm{inj}}(F) : F \in \mathcal{F})$ is a centered Gaussian process with covariance function K_{inj} given, for $F, F' \in \mathcal{F}$, by:
$$K_{\mathrm{inj}}(F, F') = \sum_{q \in V(F)} \sum_{q' \in V(F')} t\left((F \bowtie F')(q, q'), W\right) - v(F)v(F')\, t(F, W)t(F', W).$$

As a consequence, we get the central limit theorem for homomorphism densities from quantum graphs, see (4.52) and for induced homomorphism densities, see Corollary 4.15. In the Erdős-Rényi case, the one-dimensional limit distribution of induced homomorphism densities is not necessarily normal: its behaviour depends on the number of edges, two-stars and triangles in the graph F, see [159] and [118].

Notice that because $g = \mathbf{1}_{[0,D(y)]}$ is not of class $\mathcal{C}^2([0,1])$, we can not apply Theorem 4.11 (with $F = K_2$ and $k = 1$) directly to get the convergence in distribution of $\sqrt{n}(\Pi_n(y) - y)$ towards $\chi(y)$ given in Theorem 4.23. Nevertheless, the asymptotic variance can be formally obtained by computing $\sigma^{K_2,\bullet}(g)$ given in Theorem 4.11 with $g = \mathbf{1}_{[0,D(y)]}$ and $g'(z)dz = (D'(y))^{-1}\delta_{D(y)}(dz)$, with $\delta_{D(y)}(dz)$ the Dirac mass at $D(y)$. However, the proofs of Theorems 4.11 and 4.23 require different approaches.

Similarly to Theorem 4.23 and in the spirit of Theorem 4.11, it could be interesting to consider the convergence of CDF for triangles or more generally for simple finite graphs F, $V(F) = [p]$, and $\ell \in \mathcal{S}_{p,k}$:

$$\left(\frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \mathbf{1}_{\left\{ t_{\text{inj}}(F^\ell, G_n^\alpha) \leq t_x(F^\ell, W) \right\}} : x \in (0,1)^k \right),$$

where $t_x(F^\ell, W) = \mathbb{E}\left[t_{\text{inj}}\left(F^\ell, G_n^{[k]}\right) \middle| (X_1, \ldots, X_k) = x \right]$, see (4.31) and the second equality in (4.37).

4.1.4 Organization of the paper

We recall the definitions of graph homomorphisms, graphons, W-random graphs in Section 4.2. We present our main result about the almost sure convergence for the random measure $\Gamma_n^{F,\ell}$ associated to homomorphism densities of sampling partially labeled graphs from a graphon in Section 4.3.2, see Theorem 4.9. The proof is given in Section 4.5 after a preliminary result given in Section 4.4. The associated fluctuations are stated in Theorem 4.11 and proved in Section 4.6. Section 4.7 is devoted to the asymptotics for the empirical CDF of degrees Π_n, see Theorem 4.23 for the fluctuations corresponding to the almost sure convergence. After some ancillary results given in Section 4.8, we prove Theorem 4.23 in Section 4.9. We add an index of notation at the end of the paper for the reader's convenience. We postpone to the appendices some technical results on precise uniform asymptotics for the CDF of binomial distributions, see Section 4.10, and a proof of Proposition 4.28 on approximation for the CDF of multivariate binomial distributions.

4.2 Definitions

4.2.1 First notation

We denote by $|B|$ the cardinality of the set B. For $n \in \mathbb{N}^*$, we set $[n] = \{1, \ldots, n\}$. Let $\mathcal{A}$ be a non-empty set of characters, called the alphabet. A sequence $\beta = \beta_1 \ldots \beta_k$, with $\beta_i \in \mathcal{A}$ for all $1 \leq i \leq k$, is called a $\mathcal{A}$-word (or string) of length $|\beta| = k \in \mathbb{N}^*$. The word β is also identified with the vector $(\beta_1, \ldots, \beta_k)$, and for $q \in \mathcal{A}$, we write $q \in \beta$ if q belongs to $\{\beta_1, \ldots, \beta_k\}$. The concatenation of two $\mathcal{A}$-words α and β is denoted by $\alpha\,\beta$.

We now define several other operations on words. Let β be a $\mathcal{A}$-word of length $p \in \mathbb{N}^*$ and $k \in [p]$. For α a $[p]$-word of length k, we consider the $\mathcal{A}$-word β_α, defined by

$$\beta_\alpha = \beta_{\alpha_1} \ldots \beta_{\alpha_k}. \tag{4.4}$$

The word $\beta_{[k]} = \beta_1 \ldots \beta_k$ corresponds to the first k terms of β, where by convention, $[k]$ denotes the $\mathbb{N}^*$-word $1 \ldots k$. We define, for $i, j \in [p]$, the transposition word $\tau_{ij}(\beta)$ of β, obtained by exchanging the place of the ith character with the jth character in the word β: for $u \in [p]$,

$$\tau_{ij}(\beta)_u = \begin{cases} \beta_u & \text{if } u \notin \{i, j\}, \\ \beta_i & \text{if } u = j, \\ \beta_j & \text{if } u = i. \end{cases} \tag{4.5}$$

Finally, for $q \in \mathcal{A}$ and $i \in [p]$, we define the new $\mathcal{A}$-word $R_i(\beta, q)$, derived from β by substituting its ith character with q: for $u \in [p]$,

$$R_i(\beta, q)_u = \begin{cases} \beta_u & \text{if } u \neq i, \\ q & \text{if } u = i. \end{cases} \tag{4.6}$$

Let $n \in \mathbb{N}^*$ and $p \in [n]$. We define the set $\mathcal{S}_{n,p}$ of all $[n]$-words of length p such that all characters are distinct:

$$\mathcal{S}_{n,p} = \{\beta = \beta_1 \dots \beta_p : \beta_i \in [n] \text{ for all } i \in [p] \text{ and } \beta_1, \dots, \beta_p \text{ are all distinct}\}. \tag{4.7}$$

Notice that $|\mathcal{S}_{n,p}| = A_n^p = n!/(n-p)!$, and that $\mathcal{S}_{n,1} = [n]$. Moreover, for $n \in \mathbb{N}^*$, $\mathcal{S}_{n,n}$ is simply the set of all permutations of $[n]$ which will be also denoted by $\mathcal{S}_n$. With this notation, for $n \in \mathbb{N}^*$, we define the set $\mathcal{M}_n$ of all $[n]$-words with all characters distinct:

$$\mathcal{M}_n = \bigcup_{p \in [n]} \mathcal{S}_{n,p}. \tag{4.8}$$

Let $n \geq p \geq k \geq 1$ and $\ell \in \mathcal{S}_{p,k}$. For $\alpha \in \mathcal{S}_{n,k}$, we define the set $\mathcal{S}_{n,p}^{\ell,\alpha}$ of all $[n]$-words of length p such that all characters are distinct and for all $i \in [k]$, the ℓ_i-th character is equal to α_i:

$$\mathcal{S}_{n,p}^{\ell,\alpha} = \{\beta \in \mathcal{S}_{n,p} : \beta_\ell = \alpha\} \cdot \tag{4.9}$$

We have $|\mathcal{S}_{n,p}^{\ell,\alpha}| = A_{n-k}^{p-k}$. As $A_n^p = A_n^k A_{n-k}^{p-k}$, that is $|\mathcal{S}_{n,p}| = |\mathcal{S}_{n,k}|\, |\mathcal{S}_{n,p}^{\ell,\alpha}|$ for any $\alpha \in \mathcal{S}_{n,k}$, we get that for all real-valued functions f defined on $\mathcal{S}_{n,k}$:

$$\frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} f(\beta_\ell) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} f(\alpha) \cdot \tag{4.10}$$

Let $d \in \mathbb{N}^*$. For $x, y \in \mathbb{R}^d$, we denote by $\langle x, y \rangle$ the usual scalar product on $\mathbb{R}^d$ and $|x| = \sqrt{\langle x, x \rangle}$ the Euclidean norm in $\mathbb{R}^d$.

We use the convention $\prod_\emptyset = 1$.

4.2.2 Graph homomorphisms

A simple finite graph G is an ordered pair $(V(G), E(G))$ of a set $V(G)$ of $v(G)$ vertices, and a subset $E(G)$ of the collection of $\binom{v(G)}{2}$ unordered pairs of vertices. We usually shall identify $V(G)$ with $[v(G)]$. The elements of $E(G)$ are called edges and we denote by $e(G) = |E(G)|$ the number of edges in the graph G. Recall a graph G is simple when it has no self-loops, and no multiple edges between any pair of vertices. Let $\mathcal{F}$ be the set of all simple finite graphs.

Let $F, G \in \mathcal{F}$ be two simple finite graphs and set $p = v(F)$ and $n = v(G)$. A homomorphism f from F to G is an adjacency-preserving map from $V(F) = [p]$ to $V(G) = [n]$ i.e. a map from $V(F)$ to $V(G)$ such that if $\{i, j\} \in E(F)$ then $\{\mathrm{f}(i), \mathrm{f}(j)\} \in E(G)$. Let $\mathrm{Hom}(F, G)$ denote the set of homomorphisms from F to G. The homomorphism density from F to G is the normalized quantity:

$$t(F, G) = \frac{|\mathrm{Hom(F,G)}|}{n^p} \cdot \tag{4.11}$$

It is the probability that a uniform random map from $V(F)$ to $V(G)$ is a homomorphism. We have a similar definition when f is restricted to being injective. Let $\mathrm{Inj}(F, G)$ denote the set of injective homomorphisms of F into G and define its density as:

$$t_{\mathrm{inj}}(F, G) = \frac{|\mathrm{Inj}(F, G)|}{A_n^p} \cdot \tag{4.12}$$

For $\beta \in \mathcal{S}_{n,p}$, we set, with $V(F) = [p]$ and $V(G) = [n]$:

$$Y^\beta(F, G) = \prod_{\{i,j\} \in E(F)} \mathbf{1}_{\{\{\beta_i, \beta_j\} \in E(G)\}} \cdot \tag{4.13}$$

When there is no risk of confusion, we shall write Y^β for $Y^\beta(F,G)$, and thus we have:

$$t_{\text{inj}}(F,G) = \frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta\in\mathcal{S}_{n,p}} Y^\beta. \tag{4.14}$$

We recall from Lovász [138], Section 5.2.3, that:

$$|t_{\text{inj}}(F,G) - t(F,G)| \le \frac{1}{n}\binom{p}{2}. \tag{4.15}$$

In the same way, we can define homomorphism densities from partially labeled graphs. See Figure 4.1 for an injective homomorphism of partially labeled graphs. Assume $p > k \ge 1$. Let $\ell \in \mathcal{S}_{p,k}$ and $\alpha \in \mathcal{S}_{n,k}$. We define $\text{Inj}(F^\ell, G^\alpha)$ the set of injective homomorphisms f from F into G such that $\mathrm{f}(\ell_i) = \alpha_i$ for all $i \in [k]$, and its density:

$$t_{\text{inj}}(F^\ell, G^\alpha) = \frac{|\text{Inj}(F^\ell, G^\alpha)|}{A_{n-k}^{p-k}} = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta\in\mathcal{S}_{n,p}^{\ell,\alpha}} Y^\beta. \tag{4.16}$$

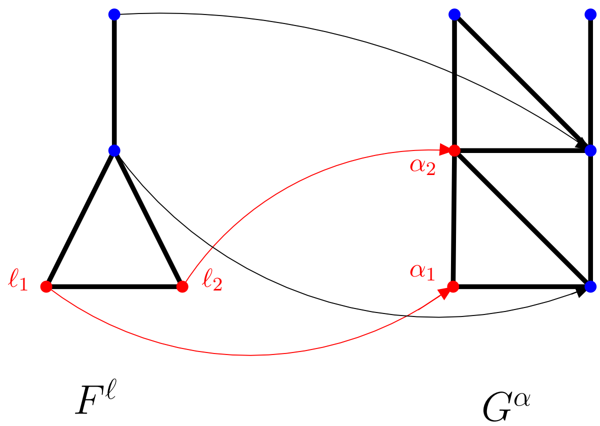

Figure 4.1 – Example of an injective homomorphism from partially labeled graphs.

Denote $\mathcal{R}_\ell(F)$ the labeled sub-graph of F with vertices $\{\ell_1, \dots, \ell_k\}$ and edges:

$$E(\mathcal{R}_\ell(F)) = \{\{i,j\} \in E(F) : i,j \in \ell\}. \tag{4.17}$$

For $\alpha \in \mathcal{S}_{n,k}$, we set:

$$\hat{Y}^\alpha(F^\ell, G^\alpha) = Y^\alpha(\mathcal{R}_\ell(F), G) = \prod_{\{i,j\}\in E(\mathcal{R}_\ell(F))} \mathbf{1}_{\{\{\alpha_i,\alpha_j\}\in E(G)\}}, \tag{4.18}$$

For $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we set $Y^\beta(F^\ell, G^\alpha) = \hat{Y}^\alpha(F^\ell, G^\alpha)\, \tilde{Y}^\beta(F^\ell, G^\alpha)$ with:

$$\tilde{Y}^\beta(F^\ell, G^\alpha) = \prod_{\{i,j\}\in E(F)\setminus E(\mathcal{R}_\ell(F))} \mathbf{1}_{\{\{\beta_i,\beta_j\}\in E(G)\}}. \tag{4.19}$$

Notice that $Y^\beta(F^\ell, G^\alpha)$ is equal to Y^β defined in (4.13) for $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$. When there is no risk of confusion, we shall write $\hat{Y}^\alpha, \tilde{Y}^\beta$ and Y^β for $\hat{Y}^\alpha(F^\ell, G^\alpha), \tilde{Y}^\beta(F^\ell, G^\alpha)$ and $Y^\beta(F^\ell, G^\alpha)$. Remark that $\hat{Y}^\alpha$ is either 0 or 1. By construction, we have:

$$t_{\text{inj}}(F^\ell, G^\alpha) = \hat{Y}^\alpha\, \tilde{t}_{\text{inj}}(F^\ell, G^\alpha) \quad \text{with} \quad \tilde{t}_{\text{inj}}(F^\ell, G^\alpha) = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta\in\mathcal{S}_{n,p}^{\ell,\alpha}} \tilde{Y}^\beta. \tag{4.20}$$

Summing (4.16) over $\alpha \in \mathcal{S}_{n,k}$, we get using (4.10) and (4.12) that:

$$\frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\text{inj}}(F^{\ell}, G^{\alpha}) = t_{\text{inj}}(F, G). \tag{4.21}$$

We can generalize this formula as follows. Let $n \geq p > k > k' \geq 1$, $\ell \in \mathcal{S}_{p,k}$, $\gamma \in \mathcal{S}_{k,k'}$ and $\alpha' \in \mathcal{S}_{n,k'}$. We easily get:

$$\frac{1}{|\mathcal{S}_{n,k}^{\gamma,\alpha'}|} \sum_{\alpha \in \mathcal{S}_{n,k}^{\gamma,\alpha'}} t_{\text{inj}}(F^{\ell}, G^{\alpha}) = t_{\text{inj}}(F^{\ell\gamma}, G^{\alpha'}). \tag{4.22}$$

Remark 4.1. Let K_2 (resp. $K_2^{\bullet}$) denote the complete graph with two vertices (resp. one of them being labeled). Let $G \in \mathcal{F}$ with n vertices. We define the degree sequence $(D_i(G) : i \in [n])$ of the graph G by, for $i \in [n]$:

$$D_i(G) = t_{\text{inj}}\left(K_2^{\bullet}, G^i\right) = \frac{1}{n-1} \sum_{j \in [n] \setminus \{i\}} \mathbf{1}_{\{\{i,j\} \in E(G)\}}. \tag{4.23}$$

Remark 4.2. Let $F \in \mathcal{F}$ be a simple finite graph with $V(F) = [p]$. Let $\ell \in \mathcal{S}_{p,k}$ for some $k \in [p]$. Assume F_0 is obtained from F by adding p' isolated vertices numbered from $p+1$ to $p+p'$, and label ℓ' of those isolated vertices so that ℓ' is a $\{p+1, \ldots, p+p'\}$-word of length say $k' = |\ell'| \leq p'$ and $\ell\ell' \in \mathcal{S}_{p+p',k+k'}$. By convention $k' = 0$ means none of the added isolated vertices is labeled. Assume $n \geq p+p'$ and let $\alpha \in \mathcal{S}_{n,k}$ and α' be a $[n]$-word such that $\alpha\alpha' \in \mathcal{S}_{n,k+k'}$. Then, it is elementary to check that:

$$t_{\text{inj}}(F_0^{\ell\ell'}, G^{\alpha\alpha'}) = t_{\text{inj}}(F^{\ell}, G^{\alpha}).$$

as well as, with δ_x the Dirac mass at x:

$$\frac{1}{|\mathcal{S}_{n,k+k'}|} \sum_{\alpha\alpha' \in \mathcal{S}_{n,k+k'}} \delta_{t_{\text{inj}}(F_0^{\ell\ell'}, G^{\alpha\alpha'})} = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \delta_{t_{\text{inj}}(F^{\ell}, G^{\alpha})}.$$

In conclusion adding isolated vertices (labeled or non labeled) does not change the homomorphism densities.

Finally, we recall an induced homomorphism from F to G is an injective homomorphism which preserves non-adjacency, that is: an injective maps f from $V(F)$ to $V(G)$ is an induced homomorphism if $\{i,j\} \in E(F)$ if and only if $\{\mathrm{f}(i), \mathrm{f}(j)\} \in E(G)$. See Figure 4.2 for an injective homomorphism which is not an induced homomorphism. Let $\text{Ind}(F,G)$ denote the set of induced homomorphisms; we denote its density by:

$$t_{\text{ind}}(F,G) = \frac{|\text{Ind}(F,G)|}{A_n^p}. \tag{4.24}$$

We recall results from [138], see Section 5.2.3., which gives relations between injective and induced homomorphism densities.

Proposition 4.3. *For $F, G \in \mathcal{F}$, two simple finite graphs, we have:*

$$t_{\text{inj}}(F,G) = \sum_{F' \geq F} t_{\text{ind}}(F',G) \quad \text{and} \quad t_{\text{ind}}(F,G) = \sum_{F' \geq F} (-1)^{e(F')-e(F)} t_{\text{inj}}(F',G), \tag{4.25}$$

where $F' \geq F$ means that $V(F) = V(F')$ and $E(F) \subset E(F')$, that is F' ranges over all simple graphs obtained from F by adding edges.

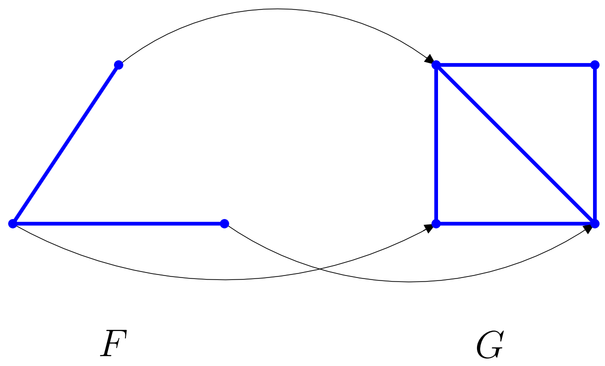

Figure 4.2 – An example of an injective homomorphism but not an induced homomorphism.

4.2.3 Graphons

A graphon is a symmetric, measurable function $W : [0,1]^2 \to [0,1]$. Denote the space of all graphons by $\mathcal{W}$. Homomorphism densities from graphs can be extended to graphons. For every simple finite graph F and every graphon $W \in \mathcal{W}$, we define

$$t(F, W) = t_{\text{inj}}(F, W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i, x_j) \prod_{k\in V(F)} dx_k \tag{4.26}$$

and

$$t_{\text{ind}}(F, W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i, x_j) \prod_{\{i,j\}\notin E(F)} (1 - W(x_i, x_j)) \prod_{k\in V(F)} dx_k. \tag{4.27}$$

A sequence of simple finite graphs $(H_n : n \in \mathbb{N}^*)$ is called convergent if the sequence $(t(F, H_n) : n \in \mathbb{N}^*)$ has a limit for every simple finite graph F. Lovász and Szegedy [139] proved that the limit of a convergent sequence of graphs can be represented as a graphon, up to a measure preserving bijection. In particular, a sequence of graphs $(G_n : n \in \mathbb{N}^*)$ is said to converge to a graphon W if for every simple finite graph F, we have

$$\lim_{n\to\infty} t(F, H_n) = t(F, W).$$

As an extension, we can define homomorphism densities from a k-labeled simple finite graph F to a graphon W which are defined by not integrating the variables corresponding to labeled vertices. Let $F \in \mathcal{F}$ be a simple finite graph, set $p = v(F)$ and identify $V(F)$ with $[p]$. Let $p \geq k \geq 1$ and $\ell \in \mathcal{S}_{n,k}$. Recall $E(\mathcal{R}_\ell(F))$ defined in (4.17). We set for $y = (y_1, \ldots, y_p) \in [0,1]^p$:

$$\tilde{Z}(y) = \prod_{\{i,j\}\in E(F)\setminus E(\mathcal{R}_\ell(F))} W(y_i, y_j) \tag{4.28}$$

and for $x = (x_1, \ldots, x_k) \in [0,1]^k$ we consider the average of $\tilde{Z}(y)$ over y restricted to $y_\ell = x$:

$$\tilde{t}_x(F^\ell, W) = \int_{[0,1]^p} \tilde{Z}(y) \prod_{m\in[p]\setminus\ell} dy_m \prod_{m'\in[k]} \delta_{x_{m'}}(dy_{\ell_{m'}}), \tag{4.29}$$

as well as the analogue of $\hat{Y}^\alpha$ for the graphon:

$$\hat{t}_x(F^\ell, W) = \prod_{\{\ell_i,\ell_j\}\in E(\mathcal{R}_\ell(F))} W(x_i, x_j) \quad \text{and} \quad \hat{t}(F^\ell, W) = \int_{[0,1]^k} \hat{t}_x(F^\ell, W)\, dx = t(\mathcal{R}_\ell(F), W). \tag{4.30}$$

Similarly to (4.20), we set for $\ell \in \mathcal{S}_{n,k}$ and $x \in [0,1]^k$:

$$t_x(F^\ell, W) = \hat{t}_x(F^\ell, W)\, \tilde{t}_x(F^\ell, W). \tag{4.31}$$

Let β and β' be $[k]$-words such that $\beta\beta' \in \mathcal{S}_k$, with $k' = |\beta'|$ and $1 \leq k' < k$. We easily get:

$$\int_{[0,1]^{k'}} t_x(F^\ell, W)\, dx_{\beta'} = t_{x_\beta}(F^{\ell_\beta}, W). \tag{4.32}$$

The result also holds for $k' = k$ with the convention $t_{x_\beta}(F^{\ell_\beta}, W) = t(F, W)$ when $\beta = \emptyset$.

Remark 4.4. The Erdős-Rényi case corresponds to $W \equiv \mathfrak{p}$ with $0 < \mathfrak{p} < 1$, and in this case we have $t(F, W) = t_x(F^\ell, W) = \mathfrak{p}^{e(F)}$ for all $x \in [0,1]^k$.

Remark 4.5. The normalized degree function D of the graphon W is defined by, for all $x \in [0,1]$:

$$D(x) = \int_0^1 W(x,y)dy. \tag{4.33}$$

We have for $W \in \mathcal{W}$ and $x \in [0,1]$:

$$t_x(K_2^\bullet, W) = \int_0^1 W(x,y)dy = D(x) \quad \text{and} \quad t(K_2, W) = \int_0^1 D(x)dx.$$

4.2.4 W-random graphs

To complete the identification of graphons as the limit object of convergent sequences, it has been proved by Lovász and Szegedy [139] that we can always find a sequence of graphs, given by a sampling method, whose limit is a given graphon function.

Let $W \in \mathcal{W}$. We can generate a W-random graph G_n with vertex set $[n]$ from the given graphon W, by first taking an independent sequence $X = (X_i : i \in \mathbb{N}^*)$ with uniform distribution on $[0,1]$, and then, given this sequence, letting $\{i,j\}$ with $i,j \in [n]$ be an edge in G_n with probability $W(X_i, X_j)$. When we need to stress the dependence on W, we shall write $G_n(W)$ for G_n. For a given sequence X, this is done independently for all pairs $(i,j) \in [n]^2$ with $i < j$.

The random graphs $G_n(W)$ thus generalize the Erdős-Rényi random graphs $G_n(\mathfrak{p})$ obtained by taking $W \equiv \mathfrak{p}$ with $0 < \mathfrak{p} < 1$ constant. (We recall that the Erdős-Rényi random graph $G_n(\mathfrak{p})$ is a random graph defined on the finite set $[n]$ of vertices whose edges occur independently with the same probability $\mathfrak{p}$, $0 < \mathfrak{p} < 1$.) Moreover, $(G_n : n \in \mathbb{N}^*)$ converges a.s. towards the graphon W, see for instance [138], Proposition 11.32.

Remark 4.6. We provide elementary computations which motivate the introduction in the previous section of $\hat{t}_x(F^\ell, W)$ and $\tilde{t}_x(F^\ell, W)$. Recall that $X_\gamma = (X_{\gamma_1}, \ldots, X_{\gamma_r})$ with γ a $\mathbb{N}^*$-word of length $|\gamma| = r$. Let $n \geq p \geq 1$ and $F \in \mathcal{F}$ with $V(F) = [p]$ and $\ell \in \mathcal{S}_{p,k}$. We set for $x = (x_1, \ldots, x_p) \in [0,1]^p$:

$$Z(x) = \prod_{\{i,j\} \in E(F)} W(x_i, x_j).$$

Let $\alpha \in \mathcal{S}_{p,k}$ and $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$. By construction, we have:

$$Z(X_\beta) = \mathbb{E}\left[Y^\beta(F, G_n) \,|\, X\right] = \mathbb{E}\left[Y^\beta(F^\ell, G_n^\alpha) \,|\, X\right].$$

By definition of $\hat{t}_x(F^\ell, W)$ and $\tilde{t}_x(F^\ell, W)$, we get:

$$\hat{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha) \,|\, X\right] = \mathbb{E}\left[Y^\alpha(\mathcal{R}_\ell(F), G_n^\alpha) \,|\, X\right] \tag{4.34}$$

$$\tilde{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\tilde{Z}(X_\beta) \,|\, X_\alpha\right] = \mathbb{E}\left[\tilde{Y}^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \tag{4.35}$$

$$t_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[Z(X_\beta) \,|\, X_\alpha\right] = \mathbb{E}\left[Y^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right]. \tag{4.36}$$

By summing (4.35) and (4.36) over $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we get, using (4.20) and (4.17), that

$$\tilde{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \quad \text{and} \quad t_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[t_{\mathrm{inj}}(F^\ell, G_n^\alpha) \,|\, X_\alpha\right]. \tag{4.37}$$

Taking the expectation in the second equality of (4.37), we deduce that:

$$t(F, W) = \int_{[0,1]^k} t_x(F^\ell, W)\, dx = \mathbb{E}\left[t_{X_\alpha}(F^\ell, W)\right] = \mathbb{E}\left[t_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right].$$

Thanks to (4.21), we recover that

$$t(F, W) = \mathbb{E}\left[t_{\mathrm{inj}}(F, G_n)\right],$$

see also [138], Proposition 11.32 or [143] Proposition A.1. We also have:

$$t(F, W) = \mathbb{E}\left[Z(X_\beta)\right] = \mathbb{E}\left[Y^\beta(F, W)\right].$$

By definition of $\hat{t}(F^\ell, W)$, we get:

$$\hat{t}(F^\ell, W) = \mathbb{E}\left[\hat{t}_{X_\alpha}(F^\ell, W)\right] = \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha)\right] = \mathbb{E}\left[Y^\alpha(\mathcal{R}_\ell(F), G_n)\right] = t(\mathcal{R}_\ell(F), W).$$

Since $\hat{Y}^\alpha(F^\ell, G_n^\alpha)$ and $\tilde{Y}^\beta(F^\ell, G_n^\alpha)$ are, conditionally on X or X_β or X_α, independent, we deduce that:

$$\begin{aligned} t_{X_\alpha}(F^\ell, W) &= \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha)\tilde{Y}^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \\ &= \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \mathbb{E}\left[\tilde{Y}^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \\ &= \hat{t}_{X_\alpha}(F^\ell, W)\; \tilde{t}_{X_\alpha}(F^\ell, W). \end{aligned}$$

This latter equality gives another interpretation of (4.31).

4.3 Asymptotics for homomorphism densities of sampling partially labeled graphs from a graphon

4.3.1 Random measures associated to a graphon

Let $d \geq 1$ and $I = [0,1]^d$. We denote by $\mathcal{B}(I)$ (resp. $\mathcal{B}^+(I)$) the set of all real-valued (resp. non negative) measurable functions defined on I. We denote by $\mathcal{C}(I)$ (resp. $\mathcal{C}_b(I)$) the set of real-valued (resp. bounded) continuous functions defined on I. For $f \in \mathcal{B}(I)$ we denote by $\|f\|_\infty$ the supremum norm of f on I. We denote by $\mathcal{C}^k(I)$ the set of real-valued functions f defined on I with continuous k-th derivative. For $f \in \mathcal{C}^1(I)$, its derivative is denoted by $\nabla f = (\nabla_1 f, \ldots, \nabla_d f)$ and we set $\|\nabla f\|_\infty = \sum_{i=1}^d \|\nabla_i f\|_\infty$.

Let $F = (F_m : 1 \leq m \leq d) \in \mathcal{F}^d$ be a finite sequence of simple finite graphs. Using Remark 4.2, if necessary, we can complete the graphs F_m with isolated vertices such that for all $m \in [d]$, we have $v(F_m) = p$ for some $p \in \mathbb{N}^*$ and consider that $V(F_m) = [p]$. We shall write $p = v(F)$. Let $\ell \in \mathcal{M}_p$ (where $\mathcal{M}_p$ is the set of all $[p]$-words with all characters distinct, given by (4.8)) and set $k = |\ell|$. We denote:

$$t_{\mathrm{inj}}(F^\ell, G_n^\alpha) = \Big(t_{\mathrm{inj}}(F_m^\ell, G_n^\alpha) : m \in [d]\Big) \in [0,1]^d,$$

and similarly for $\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)$. Let W be a graphon and $x \in [0,1]^k$. Similarly, we define $t_x(F^\ell, W)$, and $\tilde{t}_x(F^\ell, W)$, so for example:

$$t_x(F^\ell, W) = \Big(t_x(F_m^\ell, W) : m \in [d]\Big) \in [0,1]^d.$$

Notice that relabeling F_m if necessary, we get all the possible combinations of density of labeled injective homomorphism of F_m into G for all $m \in [d]$ (we could even take $\ell = [k]$).

Recall $F_m^{[\ell]}$ is the labeled sub-graph of F_m with vertices $\{\ell_1, \dots, \ell_k\}$ and set of edges $E(F_m^{[\ell]}) = \{\{i,j\} \in E(F_m) : i,j \in \ell\}$ see (4.17). For simplicity, we shall assume the following condition which states that $F_m^{[\ell]}$ does not depend on m:

$$\text{For } m, m' \in [d],\ i, i' \in \ell, \text{ we have: } \{i, i'\} \in E(F_m) \iff \{i, i'\} \in E(F_{m'}). \tag{4.38}$$

This condition can be removed when stating the main results from Section 4.3.2 at the cost of very involved notation. Therefore, we shall leave this extension to the very interested reader.

Let $G_n = G_n(W)$ be the associated W-random graphs with n vertices constructed from W and the sequence $X = (X_i : i \in \mathbb{N}^*)$ of independent uniform random variables on $[0,1]$. Under Condition (4.38), for $\alpha \in \mathcal{S}_{n,k}$ and $x \in [0,1]^k$, we have that $\hat{Y}^\alpha(F_m^\ell, G_n^\alpha)$ and $\hat{t}_x(F_m^\ell, W)$ do not depend on $m \in [d]$. We set $\hat{Y}^\alpha(F^\ell, G_n^\alpha)$ and $\hat{t}_x(F^\ell, W)$ for the common values. When there is no confusion, we write $\hat{Y}^\alpha$ for $\hat{Y}^\alpha(F^\ell, G_n^\alpha)$. In particular, we deduce from (4.20) that:

$$t_{\mathrm{inj}}(F^\ell, G_n^\alpha) = \hat{Y}^\alpha\, \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) \quad \text{with} \quad \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) = \left(\tilde{t}_{\mathrm{inj}}(F_m^\ell, G_n^\alpha) : m \in [d]\right). \tag{4.39}$$

Remark 4.7. If $|\ell| = k = 1$, then Condition (4.38) is automatically satisfied and we have by convention that $\hat{Y}^\alpha = \hat{t}_x(F^\ell, W) = 1$ for $\alpha \in \mathcal{S}_{n,k}$ and $x \in [0,1]^k$. If $d = 1$, then, Condition (4.38) is also automatically satisfied.

We define the random probability measure $\Gamma_n^{F,\ell}$ on $([0,1]^d, \mathcal{B}([0,1]^d))$ by, for $g \in \mathcal{B}^+([0,1]^d)$:

$$\begin{aligned}\Gamma_n^{F,\ell}(g) &= \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} g\left(t_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right) \\ &= \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha g\left(\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right) + (1 - \hat{Y}^\alpha) g(0),\end{aligned} \tag{4.40}$$

where we used (4.39) and the fact that $\hat{Y}^\alpha$ takes values in $\{0,1\}$ for the second equality. For $k \in \mathbb{N}^*$ and α an $\mathbb{N}^*$-word of length k, we recall the notation $X_\alpha = (X_{\alpha_1}, \dots, X_{\alpha_k})$ and $X_{[k]} = (X_1, \dots, X_k)$. Recall (4.29) and (4.30). We define the auxiliary random probability measure $\hat{\Gamma}_n^{F,\ell}$ on $[0,1]^d$ by, for $g \in \mathcal{B}^+([0,1]^d)$:

$$\hat{\Gamma}_n^{F,\ell}(g) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{t}_{X_\alpha}(F^\ell, W)\, g\left(\tilde{t}_{X_\alpha}(F^\ell, W)\right) + \left(1 - \hat{t}_{X_\alpha}(F^\ell, W)\right) g(0). \tag{4.41}$$

and the deterministic probability measure $\Gamma^{F,\ell}$, by, for all $g \in \mathcal{B}^+([0,1]^d)$:

$$\begin{aligned}\Gamma^{F,\ell}(g) &= \mathbb{E}\left[\hat{\Gamma}_n^{F,\ell}(g)\right] \\ &= \int_{[0,1]^k} \hat{t}_x(F^\ell, W)\, g\left(\tilde{t}_x(F^\ell, W)\right)\, dx + \left(1 - \hat{t}(F^\ell, W)\right) g(0).\end{aligned} \tag{4.42}$$

Remark 4.8.

(i) If $d = 1$ and $g = \mathrm{Id}$, then we have thanks to (4.21) that:

$$\Gamma_n^{F,\ell}(\mathrm{Id}) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\mathrm{inj}}(F^\ell, G_n^\alpha) = t_{\mathrm{inj}}(F, G_n)$$

and, thanks to (4.31) and (4.32):

$$\Gamma^{F,\ell}(\mathrm{Id}) = \int_{[0,1]^k} t_x(F^\ell, W)dx = t(F,W).$$

Notice that $\Gamma_n^{F,\ell}(\mathrm{Id})$ and $\Gamma^{F,\ell}(\mathrm{Id})$ do not depend on ℓ.

(ii) If $|\ell| = 1$, then according to Remark 4.7, we get:

$$\Gamma^{F,\ell}(g) = \int_{[0,1]} g\left(t_x(F^\ell, W)\right) dx.$$

4.3.2 Invariance principle and its fluctuations

We first state the invariance principle for the random probability measure $\Gamma_n^{F,\ell}$. The proof of the next theorem is given in Section 4.5.

Theorem 4.9. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $V(F) = [p]$, $\ell \in \mathcal{M}_p$. Assume that Condition (4.38) holds. Then, the sequence of random probability measures on $[0,1]^d$, $\left(\Gamma_n^{F,\ell} : n \in \mathbb{N}^*\right)$ converges a.s. for the weak topology towards $\Gamma^{F,\ell}$.*

The convergence of $\left(\Gamma_n^{F,\ell}(\mathrm{Id}) : n \in \mathbb{N}^*\right)$, with $d = 1$, can also be found in [138], see Proposition 11.32.

Remark 4.10. By Portmanteau Theorem, we have that a.s. for all bounded measurable function g on $[0,1]^d$ such that $\Gamma^{F,\ell}(\mathcal{D}_g) = 0$ where $\mathcal{D}_g$ is the set of discontinuity points of g, $\lim_{n\to\infty} \Gamma_n^{F,\ell}(g) = \Gamma^{F,\ell}(g)$.

For simplicity, consider the case $d = 1$ and $W \equiv \mathfrak{p}$ with $0 < \mathfrak{p} < 1$. Let $\hat{e}(F)$ denote the cardinality of $E(\mathcal{R}_\ell(F))$. Because $\Gamma^{F,\ell} = \mathfrak{p}^{\hat{e}(F)}\delta_{\mathfrak{p}^{e(F)-\hat{e}(F)}} + (1-\mathfrak{p}^{\hat{e}(F)})\delta_0$, with $k = |\ell|$, then if g is continuous at $\mathfrak{p}^{e(F)-\hat{e}(F)}$ and at 0, we get that a.s. $\lim_{n\to\infty} \Gamma_n^{F,\ell}(g) = \Gamma^{F,\ell}(g)$.

The next theorem, whose proof is given in Section 4.6, gives the fluctuations corresponding to the invariance principle of Theorem 4.9. Notice the speed of convergence in the invariance principle is of order $\sqrt{n}$.

For $\mu \in \mathbb{R}$ and $\sigma \geq 0$, we denote by $\mathcal{N}(\mu, \sigma^2)$ the Gaussian distribution with mean μ and variance σ^2.

Theorem 4.11. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $V(F) = [p]$, $\ell \in \mathcal{M}_p$, with $k = |\ell|$. Assume that Condition (4.38) holds. Then, for all $g \in \mathcal{C}^2([0,1]^d)$, we have the following convergence in distribution:*

$$\sqrt{n}\left(\Gamma_n^{F,\ell}(g) - \Gamma^{F,\ell}(g)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^{F,\ell}(g)^2\right),$$

with $\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(\mathcal{U}_g^{F,\ell}\right)$ and

$$\mathcal{U}_g^{F,\ell} = \sum_{i=1}^k \int_{[0,1]^k} \hat{t}_{R_i(x,U)}(F^\ell, W)\left(g(\tilde{t}_{R_i(x,U)}(F^\ell, W)) - g(0)\right) dx + \sum_{q\in[p]\backslash\ell} \int_{[0,1]^k} dx\, \langle t_{xU}(F^{\ell q}, W), \nabla g\left(\tilde{t}_x(F^\ell, W)\right)\rangle, \quad (4.43)$$

where U is a uniform random variable on $[0,1]$, and $[p]\backslash\ell = \{1,\dots,p\}\backslash\{\ell_1,\dots,\ell_k\}$.

Remark 4.12. Let U be a uniform random variable on $[0,1]$.

(i) Assume that $|\ell| = k = 1$. Using Remark 4.7, we get for $\ell \in [p]$:

$$\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(g\left(t_U(F^\ell, W)\right) + \sum_{q\in[p]\setminus\ell} \int_{[0,1]} \langle t_{xU}(F^{\ell q}, W), \nabla g\big(t_x(F^\ell, W)\big)\rangle \, dx \right). \tag{4.44}$$

(ii) Let $F \in \mathcal{F}^d$ with $p = v(F)$. Take $\ell = 1$ with $k = |\ell| = 1$. Let $a \in \mathbb{R}^d$ and consider $g(x) = \langle a, x\rangle$ for $x \in \mathbb{R}^d$. We deduce from (4.44) and (4.32) that:

$$\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(\langle a, \sum_{q=1}^{p} t_U\left(F^q, W\right)\rangle \right). \tag{4.45}$$

(iii) In the case $d = 1$, $F \in \mathcal{F}$, and $g = \mathrm{Id}$, the central limit theorem appears already in [84]. In this case, we have $\Gamma_n^{F,\ell}(\mathrm{Id}) = t_{\mathrm{inj}}(F, G_n)$, $\Gamma^{F,\ell}(\mathrm{Id}) = t(F, W)$ and, thanks to (4.45) (with $d = 1$ and $a = 1$):

$$\sigma^{F,\ell}(\mathrm{Id})^2 = \mathrm{Var}\left(\sum_{q=1}^{p} t_U\left(F^q, W\right) \right). \tag{4.46}$$

Let $F, F' \in \mathcal{F}$ be two simple finite graphs, let $i \in V(F)$ and $i' \in V(F')$. We define a new graph $(F \bowtie F')(i, i') = (F \sqcup F')/\{i \sim i'\}$ which is the disjoint union of F and F' followed by a quotient where we identify the vertex i in $V(F)$ with the vertex i' in $V(F')$, see Figure 4.3.

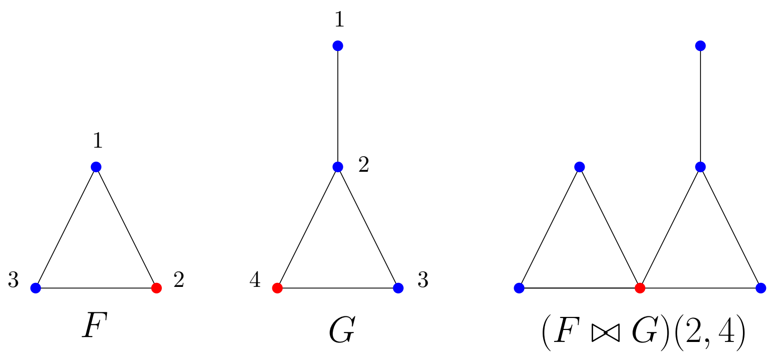

Figure 4.3 – Example of two graphs connected by two vertices.

With this notation, we have:

$$\begin{aligned}
\sigma^{F,\ell}(\mathrm{Id})^2 &= \mathbb{E}\left[\left(\sum_{q=1}^{p} t_U(F^q, W)\right)^2\right] - \mathbb{E}\left[\sum_{q=1}^{p} t_U(F^q, W)\right]^2 \\
&= \sum_{q,q'=1}^{p} \int_0^1 t_x(F^q, W) t_x(F^{q'}, W) dx - \left(\sum_{q=1}^{p} \int_0^1 t_x(F^q, W) dx\right)^2 \\
&= \sum_{q,q'=1}^{p} t\big((F \bowtie F)(q, q'), W\big) - p^2\, t(F, W)^2.
\end{aligned} \tag{4.47}$$

Thus, we recover the limiting variance given in [84].

(iv) Let $d = 1$. We consider the two degenerate cases where no vertex is labelled ($k = 0$) or all vertices are labelled ($k = p$):

(a) for $k = 0$, we apply the δ-method to (4.46), to get that

$$\sqrt{n}\,[g(t_{\text{inj}}(F, G_n)) - g(t(F, W))] \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^F(g)^2\right),$$

where

$$\sigma^F(g)^2 = g'(t(F, W))^2 \sigma^{F,\ell}(\text{Id})^2. \tag{4.48}$$

(b) for $k = p$, we have $\Gamma_n^{F,\ell}(g) = (g(1) - g(0))t_{\text{inj}}(F, G_n) + g(0)$, $\Gamma^{F,\ell}(g) = (g(1) - g(0))t(F^\ell, W)dx + g(0)$ and

$$\sigma^{F,\ell}(g)^2 = (g(1) - g(0))^2 \sigma^{F,\ell}(\text{Id})^2. \tag{4.49}$$

(v) Let $d = 1$ and $F = K_2$. We have $\Gamma_n^{K_2,\ell}(\text{Id}) = t(K_2, G_n)$ and we deduce from (4.46) that $\sigma^{K_2,\ell}(\text{Id})^2 = 4\text{Var}\ (D(U))$.

(a) If $k = 1$, then we have $\Gamma_n^{K_2^\bullet,\ell}(g) = \frac{1}{n}\sum_{i=1}^n g(D_i^{(n)})$ with $D_i^{(n)} = D_i(G_n)$ the normalized degree of i in G_n, see (4.23). We deduce from (4.44) that:

$$\sigma^{K_2^\bullet,\ell}(g)^2 = \text{Var}\ \left(g(D(U)) + \int_0^1 W(x, U)g'(D(x))\, dx\right).$$

(b) If $k = 2$, using (iv)-(2), we get from (4.49) that:

$$\sigma^{K_2^{\bullet\bullet},\ell}(g)^2 = 4(g(1) - g(0))^2 \text{Var}\ (D(U)),$$

where $K_2^{\bullet\bullet}$ denotes the complete graph K_2 with two labeled vertices.

(c) Finally, if $k = 0$, using (iv)-(1), we get from (4.48) that

$$\sigma^{K_2}(g)^2 = 4\, g'\left(\int_0^1 D(x)dx\right)^2 \text{Var}\ (D(U)).$$

(vi) Thanks to (4.15), we get that Theorems 4.9 and 4.11 also hold with t_{inj} replaced by t.

(vii) It is left to reader to check that Theorem 4.11 is degenerate, that is $\sigma^{F,\ell}(g) = 0$, in the Erdős-Rényi case, that is $W \equiv \mathfrak{p}$ for some $0 \leq \mathfrak{p} \leq 1$, or when $\int_{[0,1]^k} \hat{t}_x(F, W)\, dx = 0$ and in particular when $t(F, W) = 0$.

The next corollary gives the limiting Gaussian process for the fluctuations of $(t_{\text{inj}}(F, G_n) : F \in \mathcal{F})$.

Corollary 4.13. *We have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\,(t_{\text{inj}}(F, G_n) - t(F, W)) : F \in \mathcal{F}\right) \xrightarrow[n\to+\infty]{(fdd)} \Theta_{\text{inj}},$$

where $\Theta_{\text{inj}} = (\Theta_{\text{inj}}(F) : F \in \mathcal{F})$ is a centered Gaussian process with covariance function K_{inj} given, for $F, F' \in \mathcal{F}$, with $V(F) = [p]$ and $V(F') = [p']$, by:

$$K_{\text{inj}}(F, F') = \text{Cov}\left(\sum_{q=1}^{p} t_U(F^q, W), \sum_{q=1}^{p'} t_U\left(F'^q, W\right)\right) \tag{4.50}$$

$$= \sum_{q=1}^{p}\sum_{q'=1}^{p'} t\left((F \bowtie F')(q, q'), W\right) - pp'\ t(F, W)t(F', W). \tag{4.51}$$

Proof. We deduce from (4.45) and standard results on Gaussian vectors, the convergence, for the finite-dimensional distributions towards the Gaussian process with covariance function given by (4.50). Formula (4.51) can be derived similarly to (4.47). □

Remark 4.14. In particular, Corollary 4.13 proves a central limit theorem for quantum graphs (see Lovász [138], Section 6.1). A simple quantum graph is defined as a formal linear combination of a finite number of simple finite graphs with real coefficients. This definition makes it possible to study linear combination of homomorphism densities. For $F = (F_m : m \in [d]) \in \mathcal{F}^d$ and $a = (a_m : m \in [d]) \in \mathbb{R}^d$, we define the homomorphism density of $\mathfrak{F} = \sum_{m=1}^d a_m F_m$ for a graph G and a graphon $W \in \mathcal{W}$ as $t_{\text{inj}}(\mathfrak{F}, G) = \langle a, t_{\text{inj}}(F, G)\rangle$ and $t(\mathfrak{F}, W) = \langle a, t(F, W)\rangle$. We deduce from Corollary 4.13 the following convergence in distribution:

$$\sqrt{n}\left(t_{\text{inj}}(\mathfrak{F}, G_n) - t(\mathfrak{F}, W)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma(\mathfrak{F})^2\right), \tag{4.52}$$

where $\sigma(\mathfrak{F})^2$ is given by (4.45) or equivalently $\sigma(\mathfrak{F})^2 = \sum_{m,m'\in[d]} a_m a_{m'}\, K_{\text{inj}}(F_m, F_{m'})$.

We also have the limiting Gaussian process for the fluctuations of $(t_{\text{ind}}(F, G_n) : F \in \mathcal{F})$.

Corollary 4.15. *We have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\left(t_{\text{ind}}(F, G_n) - t_{\text{ind}}(F, W)\right) : F \in \mathcal{F}\right) \xrightarrow[n\to+\infty]{(fdd)} \Theta_{\text{ind}},$$

where $\Theta_{\text{ind}} = (\Theta_{\text{ind}}(F) : F \in \mathcal{F})$ is a centered Gaussian process with covariance function K_{ind} given, for $F_1, F_2 \in \mathcal{F}$, with $V(F_1) = [p_1]$ and $V(F_2) = [p_2]$, by:

$$\begin{aligned} &K_{\text{ind}}(F_1, F_2) \\ &= \text{Cov}\left(\sum_{F_1'\geq F_1} (-1)^{e(F_1')} \sum_{q=1}^{p_1} t_U\left((F_1')^q, W\right), \sum_{F_2'\geq F_2} (-1)^{e(F_2')} \sum_{q=1}^{p_2} t_U\left((F_2')^q, W\right)\right) \\ &= \sum_{F_1'\geq F_1,\, F_2'\geq F_2} (-1)^{e(F_1')+e(F_2')} \left(\sum_{q=1}^{p_1}\sum_{q'=1}^{p_2} t\left((F_1' \bowtie F_2')(q,q'), W\right) - p_1 p_2\; t(F_1', W) t(F_2', W)\right), \end{aligned} \tag{4.53}$$

where $F' \geq F$ means that $V(F) = V(F')$ and $E(F) \subset E(F')$, that is F' ranges over all simple graphs obtained from F by adding edges.

Proof. Notice that $t_{\text{ind}}(F, G_n)$ is a linear combination of subgraph counts by Proposition 4.3. We deduce from (4.45) and standard results on Gaussian vectors, the convergence, for the finite-dimensional distributions towards the Gaussian process with covariance function given by the first equality in (4.53) which is derived from the second formula of (4.25) and (4.45). The second equality of (4.53) can be derived similarly to (4.51). □

4.4 A preliminary result

Let $F = (F_m : m \in [d]) \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ with $|\ell| = k$ such that Condition (4.38) holds. Let $W \in \mathcal{W}$ be a graphon and $X = (X_i : i \in \mathbb{N}^*)$ be a sequence of independent uniform random variables on $[0, 1]$. Let $n \in \mathbb{N}^*$ such that $n > p$. Let $G_n = G_n(W)$ be the associated W-random graphs with vertices $[n]$, see Section 4.2.4. Recall the definitions (4.13) of $Y^\beta(F, G)$, (4.18) of $\hat{Y}^\alpha(F^\ell, G^\alpha)$ and (4.19) of $\tilde{Y}^\beta(F^\ell, G^\alpha)$ for a simple finite graph F. We set $Y^\beta = (Y^\beta(F_m, G_n) : m \in [d])$ for $\beta \in \mathcal{S}_{n,p}$ and $\tilde{Y}^\beta = (\tilde{Y}^\beta(F_m^\ell, G_n^\alpha) : m \in [d])$ as well as $Y^\alpha = Y^\alpha(F_m^\ell, G_n^\alpha)$ (which does not depend on $m \in [d]$) for $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$ and $\alpha \in \mathcal{S}_{p,k}$. Notice that for $\alpha \in \mathcal{S}_{p,k}$ and $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we have that, conditionally on X, $\hat{Y}^\alpha$ and $\tilde{Y}^\beta$ are independent, $\hat{Y}^\alpha$ is a Bernoulli random variable and:

$$Y^\beta = \hat{Y}^\alpha \tilde{Y}^\beta.$$

Recall that $t_{\text{inj}}\left(F^\ell, G_n^\alpha\right) = \left(t_{\text{inj}}\left(F_m^\ell, G_n^\alpha\right) : m \in [d]\right)$. With this notation, we get from equation (4.20) that for $\ell \in \mathcal{M}_p$ with $|\ell| = k$ and $\alpha \in \mathcal{S}_{n,k}$:

$$t_{\text{inj}}\left(F^\ell, G_n^\alpha\right) = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}} Y^\beta = \hat{Y}^\alpha \, \tilde{t}_{\text{inj}}\left(F^\ell, G_n^\alpha\right), \tag{4.54}$$

with

$$\tilde{t}_{\text{inj}}\left(F^\ell, G_n^\alpha\right) = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}} \tilde{Y}^\beta. \tag{4.55}$$

We also set $Z^\beta = \mathbb{E}[Y^\beta | X]$ and $\tilde{Z}^\beta = \mathbb{E}[\tilde{Y}^\beta | X]$. Recall (4.28). We have, for $Z^\beta = (Z_m^\beta : m \in [d])$ and $\tilde{Z}^\beta = (\tilde{Z}_m^\beta : m \in [d])$ that for $m \in [d]$:

$$Z_m^\beta = \prod_{\{i,j\} \in E(F_m)} W(X_{\beta_i}, X_{\beta_j}) \quad \text{and} \quad \tilde{Z}_m^\beta = \prod_{\{i,j\} \in \tilde{E}(F_m^\ell)} W(X_{\beta_i}, X_{\beta_j}) = \tilde{Z}_m(X_\beta).$$

We recall that $\hat{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\hat{Y}^\alpha | \, X\right] = \mathbb{E}\left[\hat{Y}^\alpha | \, X_\alpha\right]$, see (4.34), to deduce that:

$$Z^\beta = \hat{t}_{X_\alpha}(F^\ell, W) \, \tilde{Z}^\beta. \tag{4.56}$$

Lemma 4.16. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ and $W \in \mathcal{W}$ be a graphon. Let $(M_\beta : \beta \in \mathcal{S}_{n,p})$ be a sequence of $\sigma(X)$-measurable $\mathbb{R}^d$-valued random variables and $n > p$. Assume Condition (4.38) holds and that there exists a finite constant K such that for all $\beta \in \mathcal{S}_{n,p}$, we have $\mathbb{E}\left[|M_\beta|^2\right] \leq K$. Then we have:*

$$\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} \langle Y^\beta - Z^\beta, \, M_\beta \rangle\right)^2\right] \leq dK \frac{p(p-1)}{8n(n-1)}.$$

Proof. We first assume that $d = 1$. We denote by $\text{Cov}(.| \, X)$ the conditional covariance given X. We have:

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} \left(Y^\beta - Z^\beta\right) M_\beta\right)^2\right] \\
&\qquad = \frac{1}{|\mathcal{S}_{n,p}|^2} \sum_{\beta \in \mathcal{S}_{n,p}} \sum_{\gamma \in \mathcal{S}_{n,p}} \mathbb{E}\left[\mathbb{E}\left[\left(Y^\beta - \mathbb{E}[Y^\beta|X]\right)\left(Y^\gamma - \mathbb{E}[Y^\gamma|X]\right) M_\beta M_\gamma \middle| \, X\right]\right] \\
&\qquad = \frac{1}{|\mathcal{S}_{n,p}|^2} \sum_{\beta \in \mathcal{S}_{n,p}} \sum_{\gamma \in \mathcal{S}_{n,p}} \mathbb{E}\left[M_\beta M_\gamma \, \text{Cov}(Y^\beta, Y^\gamma | \, X)\right] \\
&\qquad \leq \frac{1}{|\mathcal{S}_{n,p}|^2} \sum_{\beta \in \mathcal{S}_{n,p}} \sum_{\gamma \in \mathcal{S}_{n,p}} \mathbb{E}\left[|M_\beta M_\gamma| \, |\text{Cov}(Y^\beta, Y^\gamma | \, X)|\right].
\end{aligned}$$

If the $[n]$-words β and γ have at most one character in common, that is $|\beta \bigcap \gamma| \leq 1$, then, by construction, Y^β and Y^γ are conditionally independent given X . This implies then that $\text{Cov}(Y^\beta, Y^\gamma | \, X) = 0$. If $|\beta \bigcap \gamma| > 1$, then as Y^β and Y^γ are Bernoulli random variables and we have the upper bound $|\mathrm{Cov}(Y^\beta, Y^\gamma | X)| \leq 1/4$. The number of possible choices for $\beta, \gamma \in \mathcal{S}_{n,p}$ such that $|\beta \bigcap \gamma| > 1$ is bounded from above by $A_n^p \binom{p}{2} A_{n-2}^{p-2}$. We deduce that:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,p}|}\sum_{\beta\in\mathcal{S}_{n,p}}\left(Y^\beta - Z^\beta\right)M_\beta\right)^2\right] &\leq \frac{1}{4(A_n^p)^2}\,A_n^p\binom{p}{2}A_{n-2}^{p-2}\,\mathbb{E}\left[|M_\beta M_\gamma|\right]\\ &\leq K\frac{p(p-1)}{8n(n-1)},\end{aligned}$$

where we used the Cauchy-Schwarz inequality for the last inequality to get $\mathbb{E}\left[|M_\beta M_\gamma|\right] \leq K$.

In the case $d \geq 1$, the term $\mathbb{E}\left[|M_\beta M_\gamma|\right]$ in the above inequalities has to be replaced by $\mathbb{E}\left[|M_\beta|_1\,|M_\gamma|_1\right]$, where $|\cdot|_1$ is the L^1 norm in $\mathbb{R}^d$. Then, use that $|x|_1^2 \leq d|x|^2$ and thus $\mathbb{E}\left[|M_\beta|_1\,|M_\gamma|_1\right] \leq dK$ to conclude. □

The proof of the next Lemma is similar and left to the reader (notice the next lemma is in fact Lemma 4.16 stated for $d = 1$ and the graph $\mathcal{R}_\ell(F)_m$ of the labeled vertices, see Section 4.2.2 for the definition of $\mathcal{R}_\ell(F)_m$, which thanks to condition (4.38), does not depend on $m \in [d]$). We recall that $\hat{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\hat{Y}^\alpha | X\right] = \mathbb{E}\left[\hat{Y}^\alpha | X_\alpha\right]$, see (4.34).

Lemma 4.17. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$ and $W \in \mathcal{W}$ be a graphon. Let $k \in [p]$ and $(M_\alpha : \alpha \in \mathcal{S}_{n,k})$ be a sequence of $\sigma(X)$-measurable $\mathbb{R}^d$-valued random variables and $n > p$. Assume Condition (4.38) holds and that there exists a finite constant K such that for all $\alpha \in \mathcal{S}_{n,k}$, we have $\mathbb{E}\left[|M_\alpha|^2\right] \leq K$. Then we have:*

$$\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,k}|}\sum_{\alpha\in\mathcal{S}_{n,k}}\langle\hat{Y}^\alpha - \hat{t}_{X_\alpha}(F^\ell, W),\, M_\alpha\rangle\right)^2\right] \leq dK\frac{k(k-1)}{8n(n-1)}.$$

We also state a variant of Lemma 4.16, when working conditionally on X_α for some $\alpha \in \mathcal{S}_{n,k}$.

The next result is a key ingredient in the proof of Theorems 4.9 and 4.11. Recall $\tilde{t}_x\left(F^\ell, W\right) = \left(\tilde{t}_x\left(F_m^\ell, W\right) : m \in [d]\right)$ with $\tilde{t}_x$ defined in (4.29). Notice that for all $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$:

$$\tilde{t}_{X_\alpha}\left(F^\ell, W\right) = \mathbb{E}\left[\tilde{Z}^\beta | X_\alpha\right] = \mathbb{E}\left[\tilde{t}_{\mathrm{inj}}(F^\ell, G_n) | X_\alpha\right].$$

Lemma 4.18. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ with $k = |\ell|$, $\alpha \in \mathcal{S}_{n,k}$ and $W \in \mathcal{W}$ be a graphon. Assume Condition (4.38) holds. Then, we have:*

$$\mathbb{E}\left[\left|\tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right) - \tilde{t}_{X_\alpha}\left(F^\ell, W\right)\right|^2 \Big| X_\alpha, \hat{Y}^\alpha\right] \leq d\frac{(p-k)}{4(n-k)}.$$

Proof. We consider the case $d = 1$. Recall the definition of $\tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right)$ given in (4.55). Set:

$$\mathcal{A} = \mathbb{E}\left[\frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|^2}\left(\sum_{\beta\in\mathcal{S}_{n,p}^{\ell,\alpha}}\left(\tilde{Y}^\beta - \tilde{t}_{X_\alpha}\left(F^\ell, W\right)\right)\right)^2 \Big| X_\alpha, \hat{Y}^\alpha\right].$$

Following the proof of Lemma 4.16 with $M_\beta = 1$, and using also that $\mathbb{E}\left[\tilde{Y}^\beta \,|\, X_\alpha, \hat{Y}^\alpha\right] = \tilde{t}_{X_\alpha}\left(F^\ell, W\right)$, and that $\tilde{Y}^\beta$ and $\tilde{Y}^\gamma$ are conditionally on X_α independent of $\hat{Y}^\alpha$ for $\beta, \gamma \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we get:

$$\mathcal{A} \leq \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|^2}\sum_{\beta\in\mathcal{S}_{n,p}^{\ell,\alpha}}\sum_{\gamma\in\mathcal{S}_{n,p}^{\ell,\alpha}}|\mathrm{Cov}(\tilde{Y}^\beta, \tilde{Y}^\gamma | X_\alpha)|.$$

If β and γ have no more than α in common, that is $\beta \bigcap \gamma = \alpha$, then $\tilde{Y}^\beta$ and $\tilde{Y}^\gamma$ are conditionally on X_α independent and thus $\mathrm{Cov}(\tilde{Y}^\beta, \tilde{Y}^\gamma | X) = 0$.

If $|\beta \bigcap \gamma| > |\alpha|$, then as $\tilde{Y}^\beta$ and $\tilde{Y}^\gamma$ are Bernoulli random variables, we have the upper bound $|\mathrm{Cov}(\tilde{Y}^\beta, \tilde{Y}^\gamma | X)| \leq 1/4$. The number of possible choices for $\beta, \gamma \in \mathcal{S}_{n,p}^{\ell,\alpha}$ such that $|\beta \bigcap \gamma| > |\alpha|$ is bounded from above by $A_{n-k}^{p-k}(p-k) A_{n-k-1}^{p-k-1}$. We deduce that:

$$\mathcal{A} \leq \frac{1}{4(A_{n-k}^{p-k})^2} A_{n-k}^{p-k}(p-k) A_{n-k-1}^{p-k-1} \leq \frac{(p-k)}{4(n-k)}.$$

The extension to $d \geq 1$ is direct. □

4.5 Proof of Theorem 4.9

We first state a preliminary lemma.

Lemma 4.19. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ with $k = |\ell|$, and $W \in \mathcal{W}$ be a graphon. Assume Condition (4.38) holds. Then, for all $n > k$ and $g \in \mathcal{C}^1([0,1])$, we have:*

$$\mathbb{E}\left[\left|\Gamma_n^{F,\ell}(g) - \hat{\Gamma}_n^{F,\ell}(g)\right|\right] \leq d \, \| g \|_\infty \sqrt{\frac{k(k-1)}{2n(n-1)}} + \frac{1}{2} \| \nabla g \|_\infty \sqrt{\frac{p-k}{n-k}}.$$

Proof. We first consider the case $d = 1$. Let $g \in \mathcal{C}^1([0,1])$. We first assume that $g(0) = 0$. Then, we deduce from the definition (4.40) of $\Gamma_n^{F,\ell}$ and from (4.54) and (4.55), as $\hat{Y}^\alpha \in \{0,1\}$, that:

$$\Gamma_n^{F,\ell}(g) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha \, g\left(\tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right)\right).$$

And thus, using definition (4.41) of $\hat{\Gamma}_n^{F,\ell}$, we get $\left|\Gamma_n^{F,\ell}(g) - \hat{\Gamma}_n^{F,\ell}(g)\right| \leq B_1 + B_2$ with

$$B_1 = \frac{1}{|\mathcal{S}_{n,k}|} \Big| \sum_{\alpha \in \mathcal{S}_{n,k}} \left(\hat{Y}^\alpha - \hat{t}_{X_\alpha}(F^\ell, W)\right) g\left(\tilde{t}_{X_\alpha}(F^\ell, W)\right) \Big|$$

and

$$B_2 = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha \Big| g\left(\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right) - g\left(\tilde{t}_{X_\alpha}(F^\ell, W)\right) \Big|.$$

Thanks to Lemma 4.17, we get $\mathbb{E}[B_1^2] \leq \| g \|_\infty^2 \, k(k-1)/8n(n-1)$. Thanks to Lemma 4.18, we get using Jensen's inequality that $\mathbb{E}[B_2^2] \leq \| g' \|_\infty^2 (p-k)/4(n-k)$. This gives the result when $g(0) = 0$, except there is a $1/2$ in front of $\| g \|_\infty$ in the upper bound of the Lemma. In general, use that $\Gamma_n^{F,\ell}$ and $\hat{\Gamma}_n^{F,\ell}$ are probability measures, so that $\left(\Gamma_n^{F,\ell} - \hat{\Gamma}_n^{F,\ell}\right)(g) = \left(\Gamma_n^{F,\ell} - \hat{\Gamma}_n^{F,\ell}\right)(\bar{g})$, with $\bar{g} = g - g(0)$. Then use that and $\| \bar{g} \|_\infty \leq 2 \| g \|_\infty$ to conclude. The case $d \geq 1$ is similar. □

We can now prove Theorem 4.9.

Proof of Theorem 4.9. We first consider the case $d = 1$. Let $g \in \mathcal{C}^1([0,1])$. Using Lemma 4.19 and the first Borel-Cantelli lemma, we get that a.s. $\lim_{n \to \infty} \left(\Gamma_{\phi(n)}^{F,\ell}(g) - \hat{\Gamma}_{\phi(n)}^{F,\ell}(g)\right) = 0$, with $\phi(n) = n^4$. We notice that $\hat{\Gamma}_n^{F,\ell}(g)$ is a U-statistic with kernel $\Phi_1(X_{[k]})$ where for $x \in [0,1]^k$:

$$\Phi_1(x) = \hat{t}_x \, g\left(\tilde{t}_x\right) + \left(1 - \hat{t}_x\right) g(0),$$

with $t_x = t_x(F^\ell, W)$ and the obvious variants for $\tilde{t}_x$ and $\hat{t}_x$.

Moreover, because g is uniformly bounded on $[0,1]$, we get that Var $\left(\Phi_1(X_{[k]})\right) < +\infty$ and we can apply the law of large numbers for U-statistics to obtain that a.s. $\lim_{n\to\infty} \hat{\Gamma}_n^{F,\ell}(g) = \mathbb{E}[\Phi(X_{[k]})] = \Gamma^{F,\ell}(g)$. We deduce that a.s. $\lim_{n\to\infty} \Gamma_{\phi(n)}^{F,\ell}(g) = \Gamma^{F,\ell}(g)$.

Let $n' \geq n > k$. We have $\mathcal{S}_{n,k} \subset \mathcal{S}_{n',k}$ and $\mathcal{S}_{n,p}^{\ell,\alpha} \subset \mathcal{S}_{n',p}^{\ell,\alpha}$ for $\alpha \in \mathcal{S}_{n,k}$. Recall $|\mathcal{S}_{n,k}^{\ell,\alpha}| = A_{n-k}^{p-k}$. We deduce that for $\alpha \in \mathcal{S}_{n,k}$:

$$\begin{aligned}
\left|t_{\text{inj}}(F^\ell, G_n^\alpha) - t_{\text{inj}}(F^\ell, G_{n'}^\alpha)\right| &\leq \frac{1}{A_{n'-k}^{p-k}} \left|A_{n'-k}^{p-k} - A_{n-k}^{p-k}\right| + \left|\frac{1}{A_{n'-k}^{p-k}} - \frac{1}{A_{n-k}^{p-k}}\right| A_{n-k}^{p-k} \\
&= 2\left(1 - \frac{A_{n-k}^{p-k}}{A_{n'-k}^{p-k}}\right) \\
&\leq 2\left(1 - \left(\frac{n-p}{n'-p}\right)^{p-k}\right).
\end{aligned}$$

We deduce that:

$$\begin{aligned}
\left|\Gamma_n^{F,\ell}(g) - \Gamma_{n'}^{F,\ell}(g)\right| \leq \frac{1}{A_{n'}^k} \left|A_{n'}^k - A_n^k\right| \|g\|_\infty + \left|\frac{1}{A_{n'}^k} - \frac{1}{A_n^k}\right| A_n^k \|g\|_\infty \\
+ \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \left|g\left(t_{\text{inj}}(F^\ell, G_n^\alpha)\right) - g\left(t_{\text{inj}}(F^\ell, G_{n'}^\alpha)\right)\right| \\
\leq 2\|g\|_\infty \left(1 - \left(\frac{n-k}{n'-k}\right)^k\right) + 2\|g'\|_\infty \left(1 - \left(\frac{n-p}{n'-p}\right)^{p-k}\right).
\end{aligned}$$

This implies that a.s. $\lim_{n\to\infty} \sup_{n' \in \{\phi(n),\dots,\phi(n+1)\}} \left|\Gamma_{\phi(n)}^{F,\ell}(g) - \Gamma_{n'}^{F,\ell}(g)\right| = 0$.

With the first part of the proof, we deduce that for all $g \in \mathcal{C}^1([0,1])$, a.s. $\lim_{n\to\infty} \Gamma_n^{F,\ell}(g) = \Gamma^{F,\ell}(g)$. Since there exists a convergence determining countable subset of $\mathcal{C}^1([0,1])$, we get that a.s. $\lim_{n\to\infty} \Gamma_n^{F,\ell} = \Gamma^{F,\ell}$ for the weak convergence of the measures on $[0,1]$.

The proof for $d \geq 1$ is straightforward. □

4.6 Proof of Theorem 4.11

Let $\ell \in \mathcal{M}_p$ with $k = |\ell|$. We assume Condition (4.38) holds.

Recall the random probability measures $\Gamma_n^{F,\ell}$, $\hat{\Gamma}_n^{F,\ell}$ and $\Gamma^{F,\ell}$ are defined in (4.40), (4.41) and (4.42). Let $g \in \mathcal{C}^2([0,1]^d)$. We define the U-statistic

$$U_n(g) = \frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} \Phi_2(X_\beta), \tag{4.57}$$

with kernel $\Phi_2(X_{[p]})$ given by, for $x \in [0,1]^p$:

$$\Phi_2(x) = \hat{t}_{x_\ell}\, g(\tilde{t}_{x_\ell}) + \left(1 - \hat{t}_{x_\ell}\right) g(0) + \hat{t}_{x_\ell} \langle \nabla g(\tilde{t}_{x_\ell}), \tilde{Z}(x) - \tilde{t}_{x_\ell} \rangle, \tag{4.58}$$

with $\hat{t}_y = \hat{t}_y(F^\ell, W)$, $\tilde{t}_y = \tilde{t}_y(F^\ell, W)$ for $y \in [0,1]^k$ and $\tilde{Z}(x)$ defined in (4.28). Notice that:

$$\mathbb{E}[U_n(g)] = \Gamma^{F,\ell}(g). \tag{4.59}$$

We define the random signed measure $\Lambda_n^{F,\ell} = \sqrt{n}\left[\Gamma_n^{F,\ell} - \Gamma^{F,\ell}\right]$.

Lemma 4.20. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$, with $k = |\ell|$. Assume Condition (4.38) holds. Let $g \in \mathcal{C}^2([0,1]^d)$. Then, we have that $\lim_{n\to\infty} \Lambda_n^{F,\ell}(g) - \sqrt{n}\,(U_n(g) - \mathbb{E}[U_n(g)]) = 0$ in $L^1(\mathbb{P})$.*

Proof. Recall (4.55). We write:

$$\Lambda_n^{F,\ell}(g) - \sqrt{n}\,(U_n(g) - \mathbb{E}[U_n(g)]) = R_1(n) + R_2(n) + R_3(n) \tag{4.60}$$

with

$$\begin{aligned}
R_1(n) &= \frac{\sqrt{n}}{|\mathcal{S}_{n,k}|} \sum_{\alpha\in\mathcal{S}_{n,k}} \hat{Y}^\alpha\, H_1(\alpha),\\
R_2(n) &= \frac{\sqrt{n}}{|\mathcal{S}_{n,k}|} \sum_{\alpha\in\mathcal{S}_{n,k}} \big(\hat{Y}^\alpha - \hat{t}_{X_\alpha}\big)\, H_2(\alpha),\\
R_3(n) &= \frac{\sqrt{n}}{|\mathcal{S}_{n,p}|} \sum_{\beta\in\mathcal{S}_{n,p}} \langle Y^\beta - Z^\beta,\, \nabla g(\tilde{t}_{X_{\beta_\ell}})\rangle\\
&= \frac{\sqrt{n}}{|\mathcal{S}_{n,k}|} \sum_{\alpha\in\mathcal{S}_{n,k}} \hat{Y}^\alpha\, \langle \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha), \nabla g(\tilde{t}_{X_{\beta_\ell}})\rangle - \frac{\sqrt{n}}{|\mathcal{S}_{n,p}|} \sum_{\beta\in\mathcal{S}_{n,p}} \hat{t}_{X_{\beta_\ell}} \langle \tilde{Z}(X_\beta), \nabla g(\tilde{t}_{X_{\beta_\ell}})\rangle,
\end{aligned}$$

(where we used (4.54) and (4.56) for the last equality) and

$$\begin{aligned}
H_1(\alpha) &= g\big(\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)\big) - g\big(\tilde{t}_{X_\alpha}\big) - \langle \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) - \tilde{t}_{X_\alpha},\, \nabla g\big(\tilde{t}_{X_\alpha}\big)\rangle,\\
H_2(\alpha) &= g\big(\tilde{t}_{X_\alpha}\big) - g(0) - \langle \tilde{t}_{X_\alpha}, \nabla g\big(\tilde{t}_{X_\alpha}\big)\rangle.
\end{aligned}$$

According to Lemma 4.16, we get that $\lim_{n\to\infty} R_3(n) = 0$ in $L^2(\mathbb{P})$. According to Lemma 4.17, and since $|H_2(\alpha)| \leq 2\,\|g\|_\infty + \|\nabla g\|_\infty$, we get that $\lim_{n\to\infty} R_2(n) = 0$ in $L^2(\mathbb{P})$. Since $g \in \mathcal{C}^2([0,1]^d)$, by Taylor-Lagrange inequality, we have that for all $x, y \in \mathbb{R}$,

$$|g(x) - g(y) - \langle x - y, \nabla g(x)\rangle| \leq \frac{1}{2}\,\|\nabla^2 g\|_\infty\, |x - y|^2.$$

This gives $|H_1(\alpha)| \leq \frac{1}{2}\,\|\nabla^2 g\|_\infty\, |\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) - \tilde{t}_{X_\alpha}|^2$. According to Lemma 4.18, we get that $\lim_{n\to\infty} R_1(n) = 0$ in $L^1(\mathbb{P})$. This ends the proof. □

We give a central limit theorem for the U-statistic U_n defined in (4.57).

Lemma 4.21. *Under the same hypothesis as in Lemma 4.20, we have the following convergence in distribution:*

$$\sqrt{n}\left(U_n(g) - \Gamma^{F,\ell}(g)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^{F,\ell}(g)^2\right),$$

with $\sigma^{F,\ell}(g)^2 = \mathrm{Var}\,(\mathcal{U})$ and, U being a uniform random variable on $[0,1]$:

$$\begin{aligned}
\mathcal{U} = \sum_{i=1}^k \int_{[0,1]^k} \hat{t}_{R_i(x,U)}(F^\ell, W) \left(g\big(\tilde{t}_{R_i(x,U)}(F^\ell, W)\big) - g(0)\right) dx&\\
+ \sum_{q\in[p]\setminus\ell} \int_{[0,1]^k} \langle \nabla g\big(\tilde{t}_x(F^\ell, W)\big), t_{xU}(F^{\ell q}, W)\rangle\, dx&.
\end{aligned}$$

Proof. Recall the definition of $\tau_{ij}(\beta)$ given in (4.5). The random variable $U_n(g)$ is a U-statistic with bounded kernel. Since $\mathbb{E}[U_n(g)] = \Gamma^{F,\ell}(g)$, we deduce from the central limit theorem for U-statistics, see Theorem 7.1 in [105], that $\sqrt{n}\,\big(U_n(g) - \Gamma^{F,\ell}(g)\big)$ converges in distribution towards a centered Gaussian random variable with variance $\mathrm{Var}\,(\mathcal{U}')$ and $\mathcal{U}' = \sum_{q=1}^p \mathbb{E}\,[\Phi_2(\tau_{1q}(X))|\, X_1]$, and Φ_2 given by (4.58). We first compute $\mathbb{E}\,[\Phi_2(\tau_{1q}(X))|\, X_1]$ for $q \in [p]$. We distinguish the cases $q \notin \ell$ and $q \in \ell$.

The case $q \notin \{\ell_1, \dots, \ell_k\}$

Noticing that $\tau_{1q}(X)_\ell$ does not depend on X_1, we deduce that:

$$\begin{aligned}
\mathbb{E}\left[\Phi_2(\tau_{1q}(X))\middle|\, X_1\right] &= \mathbb{E}\left[\hat{t}_{\tau_{1q}(X)_\ell}\, g\big(\tilde{t}_{\tau_{1q}(X)_\ell}\big) + \big(1-\hat{t}_{\tau_{1q}(X)_\ell}\big)\, g(0)\middle|\, X_1\right] \\
&\quad + \mathbb{E}\left[\hat{t}_{\tau_{1q}(X)_\ell}\, \langle \nabla g\big(\tilde{t}_{\tau_{1q}(X)_\ell}\big), \tilde{Z}(\tau_{1q}(X)_{[p]}) - \tilde{t}_{\tau_{1q}(X)_\ell}\rangle \middle|\, X_1\right] \\
&= C + \int_{[0,1]^k} \hat{t}_x \langle \nabla g\big(\tilde{t}_x\big), \tilde{t}_{xX_1}(F^{\ell q}, W)\rangle\, dx \\
&= C + \int_{[0,1]^k} \langle \nabla g\big(\tilde{t}_x\big), t_{xX_1}(F^{\ell q}, W)\rangle\, dx,
\end{aligned}$$

where C is a constant not depending on X_1 (which therefore will disappear when computing the variance of $\mathcal{U}'$).

The case $q \in \{\ell_1, \dots, \ell_k\}$

Let $q = \ell_i$ for some $i \in [k]$. Since $\mathbb{E}\left[\tilde{Z}(\tau_{1q}(X)_{[p]})\middle|\, \tau_{1q}(X)_\ell\right] = \tilde{t}_{\tau_{1q}(X)_\ell}$, we deduce that:

$$\begin{aligned}
\mathbb{E}\left[\Phi_2(\tau_{1q}(X))\middle|\, X_1\right] &= \mathbb{E}\left[\hat{t}_{\tau_{1q}(X)_\ell}\, g\big(\tilde{t}_{\tau_{1q}(X)_\ell}\big) + \big(1-\hat{t}_{\tau_{1q}(X)_\ell}\big)\, g(0)\middle|\, X_1\right] \\
&= g(0) + \int_{[0,1]^k} \hat{t}_{R_i(x,X_1)} \Big(g\big(\tilde{t}_{R_i(x,X_1)}\big) - g(0)\Big)\, dx.
\end{aligned}$$

Thus, we obtain that $\mathcal{U}' = \mathcal{U} + C'$ for some constant C' and:

$$\mathcal{U} = \sum_{i=1}^k \int_{[0,1]^k} \hat{t}_{R_i(x,X_1)} \Big(g\big(\tilde{t}_{R_i(x,X_1)}\big) - g(0)\Big)\, dx + \sum_{q \notin \ell} \int_{[0,1]^k} \langle \nabla g\big(\tilde{t}_x\big), t_{xX_1}(F^{\ell q}, W)\rangle\, dx.$$

This gives the result. □

Theorem 4.11 is then a direct consequence of Lemmas 4.20 and 4.21 and (4.59).

4.7 Asymptotics for the empirical degrees cumulative distribution function of the degrees

Let W be a graphon on $[0,1]$ and $n \in \mathbb{N}^*$. Recall the definition of the normalized degree function D of the graphon W given in (4.33), $D(x) = \int_{[0,1]} W(x,y)dy = t_x(K_2^\bullet, W)$. From Section 4.2.4, recall that $G_n = G_n(W)$ is the associated W-random graph with n vertices constructed from W and the sequence $X = (X_i : i \in \mathbb{N}^*)$ of independent uniform random variables on $[0,1]$. Recall the (normalized) degree sequence of a graph defined in (4.23), and set

$$D_i^{(n)} = D_i(G_n) = t_{\text{inj}}(K_2^\bullet, G_n^i)$$

the normalized degree of the vertex $i \in [n]$ in G_n. By construction of G_n, we get that conditionally on X_i , $(n-1)D_i^{(n)}$ is for $n \geq i$ a binomial random variable with parameters $(n-1, D(X_i))$. We define the empirical cumulative distribution function $\Pi_n = (\Pi_n(y) : y \in [0,1])$ of the degrees of the graph G_n by, for $y \in [0,1]$:

$$\Pi_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\left\{D_i^{(n)} \leq D(y)\right\}}. \tag{4.61}$$

Remark 4.22. If we take $g = \mathbf{1}_{[0,D(y)]}$ with $y \in [0,1]$ and $F = K_2$ in (4.3) and using the expression of $\Gamma^{F,\ell}$ given in Remark 4.8, (ii), we have that $\Pi_n(y) = \Gamma_n^{K_2,\bullet}(g)$ and $\Gamma^{K_2,\bullet}(g) = y$. If D is increasing, then $\Gamma^{K_2,\bullet}$, which is the distribution of $D(U)$, with U uniform on $[0,1]$, has no atoms and Theorem 4.9 implies that a.s. $\lim_{n\to\infty} \Pi_n(y) = y$ for all $y \in [0,1]$. Using Dini's theorem, we get that if D is increasing on $[0,1]$, then the function Π_n converges almost surely towards Id, the identity map on $[0,1]$, with respect to the uniform norm.

To get the corresponding fluctuations, we shall consider the following conditions:

$$W \in \mathcal{C}^3([0,1]^2),\ D' > 0,\ W \leq 1 - \varepsilon_0 \text{ and } D \geq \varepsilon_0 \text{ for some } \varepsilon_0 \in (0, 1/2)\,. \tag{4.62}$$

If (4.62) holds, then we have $D \in \mathcal{C}^1([0,1])$ and $D([0,1]) \subset [\varepsilon_0, 1-\varepsilon_0]$. Notice that even if (4.62) holds, the set $\{W = 0\}$ might have positive Lebesgue measure; but the regularity conditions on W rules out bipartite graphons (but not tripartite graphons).

Theorem 4.23. *Assume that W satisfies condition (4.62). Then we have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\,(\Pi_n(y) - y) : y \in (0,1)\right) \xrightarrow[n\to+\infty]{(fdd)} \chi,$$

where $\chi = (\chi_y : y \in (0,1))$ is a centered Gaussian process defined, for all $y \in (0,1)$ by:

$$\chi_y = \int_0^1 (\rho(y,u) - \bar{\rho}(y))dB_u, \tag{4.63}$$

with $B = (B_u, u \geq 0)$ a standard Brownian motion, and $(\rho(y,u) : u \in [0,1])$ and $\bar{\rho}(y)$ defined for $y \in (0,1)$ by:

$$\rho(y,u) = \mathbf{1}_{[0,y]}(u) - \frac{W(y,u)}{D'(y)} \quad \text{and} \quad \bar{\rho}(y) = \int_0^1 \rho(y,u)du.$$

Remark 4.24. The covariance kernel of the Gaussian process χ can also be written as $\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_3$, where for $y, z \in (0,1)$:

$$\Sigma_1(y,z) = y \wedge z - yz, \tag{4.64}$$

$$\Sigma_2(y,z) = \frac{1}{D'(y)D'(z)} \left(\int_0^1 W(y,x)W(z,x)dx - D(y)D(z) \right), \tag{4.65}$$

$$\Sigma_3(y,z) = \frac{1}{D'(y)} \left(D(y)z - \int_0^z W(y,x)dx \right) + \frac{1}{D'(z)} \left(D(z)y - \int_0^y W(z,x)dx \right). \tag{4.66}$$

Thus, for $y \in (0,1)$ the variance of $\chi(y)$ is:

$$\Sigma(y,y) = y(1-y) + \frac{1}{D'(y)^2} \left(\int_0^1 W(y,x)^2 dx - D(y)^2 \right) + \frac{2}{D'(y)} \left(D(y)y - \int_0^y W(y,x)dx \right).$$

Remark 4.25. We conjecture that the convergence of Theorem 4.23 holds for the process in the Skorokhod space. However, the techniques used to prove this theorem are not strong enough to get such a result.

4.8 Preliminary results for the empirical cdf of the degrees

4.8.1 Estimates for the first moment of the empirical cdf

Recall $X = (X_n : n \in \mathbb{N}^*)$ is a sequence of independent random variables uniformly distributed on $[0,1]$ used to construct the sequence of W-random graphs $(G_n : n \in \mathbb{N}^*)$. Recall $\Pi_n(y)$ is given in (4.61).

For all $y \in (0,1)$, we set $c_n(y) = \mathbb{E}[\Pi_{n+1}(y)]$ that is

$$c_n(y) = \mathbb{P}\left(D_1^{(n+1)} \leq D(y)\right), \tag{4.67}$$

where, conditionally on X, $D_1^{(n+1)}$ is a binomial random variable with parameter $(n, D(X_1))$. We set:

$$\sigma^2_{(x)} = x(1-x) \quad \text{for } x \in [0,1], \tag{4.68}$$

and with $\lceil x \rceil$ the unique integer such that $\lceil x \rceil - 1 < x \leq \lceil x \rceil$,

$$S(x) = \lceil x \rceil - x - \frac{1}{2} \quad \text{for } x \in \mathbb{R}. \tag{4.69}$$

The next proposition gives precise asymptotics of c_n.

Proposition 4.26. *Assume that W satisfies condition (4.62). For all $y \in (0,1)$, there exists a constant $C > 0$ such that for all $n \in \mathbb{N}^*$, we have with $d = D(y)$,*

$$n\left(c_n(y) - y\right) = -\frac{D''(y)}{D'(y)^3}\frac{\sigma^2_{(d)}}{2} + \frac{1}{D'(y)}\left(\frac{1-2d}{2} + S(nd)\right) + R_n^{4.26},$$

with

$$\left|R_n^{4.26}\right| \leq C\, n^{-\frac{1}{4}}.$$

In particular, because $|S(x)| \leq \frac{1}{2}$, for all $x \in \mathbb{R}$, we have that for all $y \in (0,1)$:

$$c_n(y) - y = O\left(n^{-1}\right). \tag{4.70}$$

Proof. For $n \in \mathbb{N}^*$, $d \in [0,1]$, $\delta \in \mathbb{R}$ and $\mathfrak{p} \in (0,1)$, we consider the CDF:

$$\mathcal{H}_{n,d,\delta}(\mathfrak{p}) = \mathbb{P}\left(X \leq nd + \delta\right),$$

where X a binomial random variable with parameters $(n, \mathfrak{p})$. Let $y \in (0,1)$. We have:

$$c_n(y) - y = \int_0^1 \left(\mathcal{H}_{n,d,0}(D(x)) - \mathbf{1}_{\{x \leq y\}}\right) dx.$$

By Proposition 4.40 applied with $G(x) = 1$ and $\delta = 0$, we obtain that:

$$n\int_0^1 \left(\mathcal{H}_{n,d,0}\left(D(x)\right) - \mathbf{1}_{\{x \leq y\}}\right) dx = -\frac{D''(y)}{D'(y)^3}\frac{\sigma^2_{(d)}}{2} + \frac{1}{D'(y)}\left(\frac{1-2d}{2} + S(nd)\right) + R_n^{4.26},$$

with $R_n^{4.26} = R_n^{4.40}(1)$ and $\left|R_n^{4.26}\right| \leq Cn^{-\frac{1}{4}}$. □

For $y \in (0,1)$ and $u \in [0,1]$, we set, with $d = D(y)$,

$$H_n(y,u) = n\left(\mathbb{E}\left[\mathbf{1}_{\left\{D_1^{(n+1)} \leq d\right\}}\middle| X_2 = u\right] - c_n(y)\right) \tag{4.71}$$

and

$$H_n^\star(y,u) = \mathbb{E}\left[\mathbf{1}_{\left\{D_1^{(n+1)} \leq d\right\}}\middle| X_1 = u\right] - c_n(y). \tag{4.72}$$

Proposition 4.27. *Assume that W satisfies condition (4.62). For all $y \in (0,1)$, there exists a positive constant C such that for all $n \geq 2$ and $u \in [0,1]$, we have with $d = D(y)$:*

$$H_n(y,u) = \frac{1}{D'(y)}\left(d - W(y,u)\right) + R_n^{4.27}(u),$$

with

$$\left|R_n^{4.27}(u)\right| \leq C\, n^{-\frac{1}{4}}. \tag{4.73}$$

For all $y \in (0,1)$ and $u \in [0,1]$ such that $u \neq y$, we have

$$|H_n^\star(y,u)| \leq 1 \quad \text{for all } n \geq 2, \text{ and} \quad \lim_{n\to\infty} H_n^\star(y,u) = \mathbf{1}_{\{u \leq y\}} - y. \tag{4.74}$$

Proof. In what follows, C denotes a positive constant which depends on ε_0, W and $y \in (0,1)$, and it may vary from line to line. Recall that $X_{[2]} = (X_1, X_2)$. We define the function φ_n by:

$$\varphi_n(x,u) = \mathbb{P}\left(D_1^{(n+1)} \leq d \,\middle|\, X_{[2]} = (x,u)\right) - \mathbf{1}_{\{x \leq y\}} \quad \text{for } x, u \in [0,1]. \tag{4.75}$$

Then we have for $u \in [0,1]$:

$$H_n(y,u) = n\mathbb{E}\left[\varphi_n(X_1,u)\right] - n(c_n(y) - y). \tag{4.76}$$

Conditionally on $\{X_{[2]} = (x,u)\}$, $D_1^{(n+1)}$ is distributed as $Y_{12} + \tilde{B}^{(n)}$, where Y_{12} and $\tilde{B}$ are independent, Y_{12} is Bernoulli $W(x,u)$ and $\tilde{B}$ is binomial with parameter $(n-1, D(x))$. Thus, we have:

$$\begin{aligned}\varphi_n(x,u) &= \mathbb{P}\left(Y_{12} + \tilde{B} \leq nd\right) - \mathbf{1}_{\{x \leq y\}} \\ &= W(x,u)\left[\mathbb{P}\left(\tilde{B} \leq nd - 1\right) - \mathbf{1}_{\{x \leq y\}}\right] + (1 - W(x,u))\left[\mathbb{P}\left(\tilde{B} \leq nd\right) - \mathbf{1}_{\{x \leq y\}}\right] \\ &= W(x,u)\left[\mathcal{H}_{n-1,d,d-1}(D(x)) - \mathbf{1}_{\{x \leq y\}}\right] + (1 - W(x,u))\left[\mathcal{H}_{n-1,d,d}(D(x)) - \mathbf{1}_{\{x \leq y\}}\right].\end{aligned} \tag{4.77}$$

Let $W_1(x,u)$ denote $\partial W(x,u)/\partial x$. We apply Proposition 4.40 with $G(x) = W(x,u)$, $\delta = d-1$ and n replaced by $n-1$ to get that:

$$\begin{aligned}(n-1)\mathbb{E}&\left[W(X_1,u)\left[\mathcal{H}_{n-1,d,d-1}(D(X_1)) - \mathbf{1}_{\{X_1 \leq y\}}\right]\right] \\ &= \frac{\sigma_{(d)}^2}{2D'(y)^2}\left[W_1(y,u) - \frac{W(y,u)D''(y)}{D'(y)}\right] \\ &\quad + \frac{W(y,u)}{D'(y)}\left(-\frac{1}{2} + S(nd-1)\right) + R_{n-1}^{4.40}(W(.,u)),\end{aligned} \tag{4.78}$$

and with $G(x) = 1 - W(x,u)$, $\delta = d$ and n replaced by $n-1$, to get that:

$$\begin{aligned}(n-1)\mathbb{E}&\left[(1 - W(X_1,u))\left[\mathcal{H}_{n-1,d,d}(D(X_1)) - \mathbf{1}_{\{X_1 \leq y\}}\right]\right] \\ &= \frac{\sigma_{(d)}^2}{2D'(y)^2}\left[-W_1(y,u) - \frac{(1 - W(y,u))D''(y)}{D'(y)}\right] \\ &\quad + \frac{1 - W(y,u)}{D'(y)}\left(\frac{1}{2} + S(nd)\right) + R_{n-1}^{4.40}(1 - W(.,u)).\end{aligned} \tag{4.79}$$

By equations (4.77), (4.78) and (4.79) and since $S(nd-1) = S(nd)$, we get that:

$$(n-1)\mathbb{E}\left[\varphi_n(X_1,u)\right] = -\frac{\sigma_{(d)}^2}{2}\frac{D''(y)}{D'(y)^3} + \frac{1}{D'(y)}\left(\frac{1}{2} - W(y,u) + S(nd)\right) + R_n^{(1)}(u),$$

where $R_n^{(1)}(u) = R_{n-1}^{4.40}(W(.,u)) + R_{n-1}^{4.40}(1 - W(.,u))$. Because W satisfies condition (4.62), we deduce from (4.122) that $\left|R_n^{(1)}(u)\right| \leq Cn^{-1/4}$ for some finite constant C which does not depend on n and $u \in [0,1]$. Using Proposition 4.26, we get that

$$(n-1)\mathbb{E}\left[\varphi_n(X_1,u)\right] - (n-1)(c_n(y) - y) = \frac{d - W(y,u)}{D'(y)} + R_n^{(2)}(u), \tag{4.80}$$

where $R_n^{(2)}(u) = R_n^{(1)}(u) + R_n^{4.26} + (c_n(y) - y)$ and $\left|R_n^{(2)}(u)\right| \leq Cn^{-1/4}$ because of (4.70). By equations (4.76) and (4.80), we deduce that:

$$H_n(y,u) = \frac{n}{n-1}\frac{d - W(y,u)}{D'(y)} + \frac{n}{n-1}R_n^{(2)}(u) = \frac{d - W(y,u)}{D'(y)} + R_n^{4.27}(u),$$

with $\left|R_n^{4.27}(y,u)\right| \leq Cn^{-\frac{1}{4}}$. This gives (4.73).

For the second assertion (4.74), we notice that

$$H_n^\star(y,u) = \mathcal{H}_{n,d,0}(D(u)) - c_n(y),$$

with $\mathcal{H}_{n,d,0}(D(u)) \in [0,1]$ and $c_n(y) \in [0,1]$. By the strong law of large numbers, we have for $u \neq y$:

$$\lim_{n\to\infty} \mathcal{H}_{n,d,0}(D(u)) = \mathbf{1}_{\{u\leq y\}}.$$

Using (4.70), we get the expected result. □

4.8.2 Estimates for the second moment of the empirical cdf

For $\mathrm{y} = (y_1, y_2) \in [0,1]^2$, let $M(\mathrm{y})$ be the covariance matrix of a pair (Y_1, Y_2) of Bernoulli random variables such that $\mathbb{P}(Y_i = 1) = D(y_i)$ for $i \in \{1,2\}$ and $\mathbb{P}(Y_1 = Y_2 = 1) = \int_{[0,1]} W(y_1,z)W(y_2,z)\,dz$.

Let $\mathcal{K}$ be the set of all convex sets in $\mathbb{R}^2$. For $K \in \mathcal{K}$, we define its sum with a vector x in $\mathbb{R}^2$ as

$$K + \mathrm{x} = \{k + \mathrm{x} : k \in K\}$$

and its product with a real matrix A of size 2×2 as

$$AK = \{Ak : k \in K\}.$$

Recall that for $\mathrm{x} \in \mathbb{R}^2$, $|\mathrm{x}|$ is the Euclidean norm of x in $\mathbb{R}^2$. Recall $X_{[2]} = (X_1, X_2)$. We define $\mathcal{D}^{(n+1)} = (\mathcal{D}_1^{(n+1)}, \mathcal{D}_2^{(n+1)})$, where for $i \in \{1,2\}$, $\mathcal{D}_i^{(n+1)}$ is the number of edges from the vertex i to the vertices $\{k, 3 \leq k \leq n+1\}$ of G_{n+1}; it is equal to $nD_i^{(n+1)}$ if the edge $\{1,2\}$ does not belong to G_{n+1} and to $nD_i^{(n+1)} - 1$ otherwise. The proof of the next proposition is postponed to section 4.11.

Proposition 4.28. *Assume that W satisfies condition (4.62). There exists a finite constant C_0 such that for all $\mathrm{x} = (x_1, x_2) \in [0,1]^2$ with $x_1 \neq x_2$, we get for all $n \geq 2$:*

$$\sup_{K\in\mathcal{K}} \left| \mathbb{P}\left(\mathcal{D}^{n+1} \in K \middle| X_{[2]} = \mathrm{x}\right) - \mathbb{P}\left(Z \in \frac{M(\mathrm{x})^{-\frac{1}{2}}}{\sqrt{n-1}}(K - \mu(\mathrm{x}))\right)\right| \leq \frac{C_0}{\sqrt{n}},$$

where $\mu(\mathrm{x}) = (n-1)(D(x_1), D(x_2))$ and Z is a standard 2-dimensional Gaussian vector.

For $y_1, y_2 \in (0,1)$, with $d_1 = D(y_1)$ and $d_2 = D(y_2)$, we set with $\delta \in \mathbb{R}$ and $\mathrm{x} = (x_1, x_2) \in [0,1]^2$ such that $x_1 \neq x_2$:

$$\Psi_{n,\delta}(\mathrm{x}) = \mathbb{E}\left[\prod_{i\in\{1,2\}} \left(\mathbf{1}_{\{\mathcal{D}_i^{(n+1)} \leq nd_i + \delta\}} - \mathbf{1}_{\{X_i \leq y_i\}}\right) \middle| X_{[2]} = \mathrm{x}\right].$$

Recall Σ_2 defined in (4.65) and that $X_{[2]} = (X_1, X_2)$.

Lemma 4.29. *Assume that W satisfies condition (4.62). For all $\mathrm{y} = (y_1, y_2) \in (0,1)^2$, $\delta \in [-1,0]$ and $G \in \mathcal{C}^1([0,1]^2)$, we have:*

$$\lim_{n\to\infty} n\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}\left(X_{[2]}\right)\right] = G(\mathrm{y})\Sigma_2(\mathrm{y}).$$

Proof. Let $A = 4\sqrt{\log(n-1)}$. For $n \geq 2$ and $\delta \in [-1,0]$, we set:

$$\Psi_{n,\delta}^{(1)}(\mathrm{x}) = \Psi_{n,\delta}(\mathrm{x}) \prod_{i=1}^{2} \mathbf{1}_{\{\sqrt{n-1}|D(x_i) - d_i| \leq A\}} \quad \text{and} \quad \Psi_{n,\delta}^{(2)}(\mathrm{x}) = \Psi_{n,\delta}(\mathrm{x}) - \Psi_{n,\delta}^{(1)}(\mathrm{x}).$$

Then we have

$$\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}\left(X_{[2]}\right)\right] = \sum_{i\in\{1,2\}} \mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}^{(i)}\left(X_{[2]}\right)\right]. \tag{4.81}$$

Study of $\mathbb{E}\left[G\left(X_{[2]}\right)\Psi^{(2)}_{n,\delta}\left(X_{[2]}\right)\right]$

Recall that for $i \in \{1,2\}$, conditionally on $X_i = x_i$, $\mathcal{D}_i^{(n+1)}$ is distributed as a Bernoulli random variable with parameter $(n-1, d_i)$. We get that:

$$\begin{aligned}\left|\Psi^{(2)}_{n,\delta}(\mathrm{x})\right| &\leq 2 \sum_{i\in\{1,2\}} \mathbb{E}\left[\left|\mathbf{1}_{\{\mathcal{D}_i^{(n+1)}\leq nd_i+\delta\}} - \mathbf{1}_{\{X_i\leq y_i\}}\right| \mathbf{1}_{\{\sqrt{n-1}\,|D(x_i)-d_i|\geq A\}}\middle| X_{[2]} = \mathrm{x}\right] \\ &= 2\sum_{i\in\{1,2\}} \left|\mathcal{H}_{n-1,d_i,\delta+d_i}(x_i) - \mathbf{1}_{\{x_i\leq y_i\}}\right| \mathbf{1}_{\{\sqrt{n-1}\,|D(x_i)-d_i|\geq A\}}.\end{aligned}$$

By Lemma 4.39 (with n replaced by $n-1$), we deduce that:

$$\lim_{n\to\infty} n\mathbb{E}\left[G\left(X_{[2]}\right)\Psi^{(2)}_{n,\delta}\left(X_{[2]}\right)\right] = 0. \tag{4.82}$$

Study of $\mathbb{E}\left[G\left(X_{[2]}\right)\Psi^{(1)}_{n,\delta}\left(X_{[2]}\right)\right]$

This part is more delicate. For $\mathfrak{z} = (z_1, z_2) \in [0,1]^2$, set $H(\mathfrak{z}) = \frac{G(\mathfrak{z})}{D'(z_1)D'(z_2)}$ and $\mathfrak{t}_n(\mathfrak{z}) = (t_n(z_1), t_n(z_2))$:

$$t_n(z_i) = D^{-1}\left(d_i + \frac{z_i}{\sqrt{n-1}}\right) \quad \text{for } i \in \{1,2\}.$$

Using the change of variable $z_i = \sqrt{n-1}(D(x_i) - d_i)$ for $i \in \{1,2\}$ with $\mathrm{x} = (x_1, x_2)$, we get:

$$\begin{aligned}(n-1)\mathbb{E}\left[G\left(X_{[2]}\right)\Psi^{(1)}_{n,\delta}\left(X_{[2]}\right)\right] &= (n-1)\int_{[0,1]^2} G(\mathrm{x})\Psi_{n,\delta}(\mathrm{x}) \prod_{i\in\{1,2\}} \mathbf{1}_{\{\sqrt{n-1}|D(x_i)-d_i|\leq A\}} d\mathrm{x} \\ &= \int_{[-A,A]^2} H\left(\mathfrak{t}_n(\mathfrak{z})\right)\Psi_{n,\delta}\left(\mathfrak{t}_n(\mathfrak{z})\right) d\mathfrak{z}. \end{aligned}\tag{4.83}$$

Notice that:

$$\Psi_{n,\delta}\left(\mathfrak{t}_n(\mathfrak{z})\right) = \mathbb{E}\left[\prod_{i\in\{1,2\}}\left(\mathbf{1}_{\{\mathcal{D}_i^{(n+1)}\leq nd_i+\delta\}} - \mathbf{1}_{\{z_i\leq 0\}}\right)\middle| X_{[2]} = \mathfrak{t}_n(\mathfrak{z})\right].$$

Set $\tilde{\delta} = (\delta, \delta)$. We define the sets for $\mathfrak{z}$ and $\mathcal{D}^{(n+1)}$:

$$\begin{aligned} I^{(1)} &= [0,A]^2 &\text{and}\quad \tilde{C}_n^{(1)} &= \tilde{\delta} + n\left(-\infty, d_1\right] \times \left(-\infty, d_2\right], \\ I^{(2)} &= [0,A]\times[-A,0) &\text{and}\quad \tilde{C}_n^{(2)} &= \tilde{\delta} + n\left(-\infty, d_1\right] \times \left(d_2, +\infty\right), \\ I^{(3)} &= [-A,0)\times[0,A] &\text{and}\quad \tilde{C}_n^{(3)} &= \tilde{\delta} + n\left(d_1, +\infty\right) \times \left(-\infty, d_2\right], \\ I^{(4)} &= [-A,0)^2 &\text{and}\quad \tilde{C}_n^{(4)} &= \tilde{\delta} + n\left(d_1, +\infty\right) \times \left(d_2, +\infty\right). \end{aligned}$$

For $1 \leq i \leq 4$, we set:

$$Q_n^{(i)}(\mathfrak{z}) = \mathbb{P}\left(\mathcal{D}^{(n+1)} \in \tilde{C}_n^{(i)} \middle| X_{[2]} = \mathfrak{t}_n(\mathfrak{z})\right) \quad \text{and} \quad \Delta_n^{(i)} = \int_{I^{(i)}} H\left(\mathfrak{t}_n(\mathfrak{z})\right) Q_n^{(i)}(\mathfrak{z})\, d\mathfrak{z}.$$

By construction, we have:

$$(n-1)\mathbb{E}\left[G\left(X_{[2]}\right)\Psi^{(1)}_{n,\delta}\left(X_{[2]}\right)\right] = \sum_{i=1}^{4} \Delta_n^{(i)}. \tag{4.84}$$

We now study $\Delta_n^{(1)}$. By Proposition 4.28, we get that

$$\Delta_n^{(1)} = \int_{[0,A]^2} H(\mathfrak{t}_n(\mathfrak{z}))\, \mathbb{P}\left(Z \in \frac{M(\mathfrak{t}_n(\mathfrak{z}))^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right) d\mathfrak{z} + R_n^{(1)}$$

where $\mu(\mathrm{x}) = (n-1)(D(x_1), D(x_2))$ and $|R_n^{(1)}| \leq \| H \|_\infty 8C_0\sqrt{\log(n)/n}$ so that $\lim_{n\to\infty} R_n^{(1)} = 0$. Set $\tilde{d} = (d_1, d_2)$. Since $\mathfrak{t}_n(\mathfrak{z})$ converges towards y, we get:

$$\lim_{n\to\infty} H(\mathfrak{t}_n(\mathfrak{z})) = H(\mathrm{y}) \quad \text{and} \quad \lim_{n\to\infty} M(\mathfrak{t}_n(\mathfrak{z})) = M(\mathrm{y})$$

and, with $J(\mathfrak{z}) = (-\infty, -z_1] \times (-\infty, -z_2]$,

$$\frac{1}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right) = (n-1)^{-1/2}(\tilde{\delta} + \tilde{d}) + J(\mathfrak{z}). \tag{4.85}$$

Since $\lim_{n\to\infty} M(\mathfrak{t}_n(\mathfrak{z})) = M(\mathrm{y})$ and $M(\mathrm{y})$ is positive definite, we deduce that dx-a.e.:

$$\lim_{n\to\infty} \mathbf{1}_{\frac{M(\mathfrak{t}_n(\mathfrak{z}))^{-1/2}}{\sqrt{n-1}}\left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)}(\mathrm{x}) = \mathbf{1}_{M(\mathrm{y})^{-1/2}J(\mathfrak{z})}(\mathrm{x})$$

and thus (by dominated convergence):

$$\lim_{n\to\infty} \mathbb{P}\left(Z \in \frac{M(\mathfrak{t}_n(\mathfrak{z}))^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right) = \mathbb{P}\left(M(\mathrm{y})^{1/2} Z \in J(\mathfrak{z})\right).$$

For $\mathfrak{z} \in [2, +\infty)^2$ and $n \geq 2$, we have:

$$\begin{aligned}\left\{Z \in \frac{M(\mathfrak{t}_n(\mathfrak{z}))^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right\} &\subset \left\{2M(\mathfrak{t}_n(\mathfrak{z}))^{1/2} Z \in J(\mathfrak{z})\right\} \\ &\subset \left\{2^{3/2}|Z| \geq |\mathfrak{z}|\right\},\end{aligned}$$

where we used (4.85) and that for all $i \in \{1, 2\}$, $|(n-1)^{-1/2}(\delta+d_i)| \leq z_i/2$ for the first inclusion and for the second that $|M\mathrm{x}| \leq \sqrt{2}\, \| M \|_\infty |\mathrm{x}|$, $\| M^{1/2} \|_\infty \leq \sqrt{2} \| M \|_\infty^{1/2}$ and $\| M(\mathrm{x}) \|_\infty \leq 1/2$ for all $\mathrm{x} \in [0, 1]^2$ so that $|M(\mathfrak{t}_n(\mathfrak{z}))^{1/2} Z| \leq \sqrt{2}\, |Z|$. Since $\int_{\mathbb{R}} \mathbb{P}(2^{3/2}|Z| \geq |\mathfrak{z}|)\, d\mathfrak{z}$ is finite and H is bounded, we deduce from dominated convergence that:

$$\begin{aligned}\lim_{n\to\infty} \int_{[0,A]^2} H(\mathfrak{t}_n(\mathfrak{z}))\, \mathbb{P}\left(Z \in \frac{M(\mathfrak{t}_n(\mathfrak{z}))^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right) d\mathfrak{z} \\ = H(\mathrm{y}) \int_{[0,+\infty)^2} \mathbb{P}\left(M(\mathrm{y})^{1/2} Z \in J(\mathfrak{z})\right) d\mathfrak{z}.\end{aligned}$$

Recall $x^+ = \max(0, x)$ and $x^- = \max(0, -x)$ denote the positive and negative part of $x \in \mathbb{R}$. With $\tilde{Z} = M(\mathrm{y})^{1/2} Z = (\tilde{Z}_1, \tilde{Z}_2)$, we get:

$$\int_{[0,+\infty)^2} \mathbb{P}\left(M(\mathrm{y})^{1/2} Z \in J(\mathfrak{z})\right) d\mathfrak{z} = \mathbb{E}\left[\tilde{Z}_1^- \tilde{Z}_2^-\right]$$

and thus

$$\lim_{n\to\infty} \Delta_n^{(1)} = H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^- \tilde{Z}_2^-\right].$$

Similarly, we obtain:

$$\begin{aligned}\lim_{n\to\infty} \Delta_n^{(2)} &= -H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^- \tilde{Z}_2^+\right] \\ \lim_{n\to\infty} \Delta_n^{(3)} &= -H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^+ \tilde{Z}_2^-\right] \\ \lim_{n\to\infty} \Delta_n^{(4)} &= H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^+ \tilde{Z}_2^+\right].\end{aligned}$$

Using the definition of $\Sigma_2(\mathrm{y})$ and $M(\mathrm{y})$, notice that $D'(y_1)D'(y_2)\Sigma_2(\mathrm{y})$ is the covariance of $\tilde{Z}_1$ and $\tilde{Z}_2$. Thus, we obtain:

$$\lim_{n\to\infty}\sum_{i=1}^{4}\Delta_n^{(i)} = H(\mathrm{y})\mathbb{E}\left[\left(\tilde{Z}_1^+ - \tilde{Z}_1^-\right)\left(\tilde{Z}_2^+ - \tilde{Z}_2^-\right)\right] = H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1\tilde{Z}_2\right] = G(\mathrm{y})\Sigma_2(\mathrm{y}). \quad (4.86)$$

Conclusion

Use (4.81), (4.82), (4.84) and (4.86) to get the result. □

The next proposition, the main result of this section, is a consequence of Lemma 4.29.

Proposition 4.30. *Assume that W satisfies condition (4.62). For all $y_1, y_2 \in (0,1)$, we have with $d_1 = D(y_1)$ and $d_2 = D(y_2)$:*

$$\lim_{n\to\infty} n\,\mathbb{E}\left[\left(\mathbf{1}_{\left\{D_1^{(n+1)}\leq d_1\right\}} - \mathbf{1}_{\{X_1\leq y_1\}}\right)\left(\mathbf{1}_{\left\{D_2^{(n+1)}\leq d_2\right\}} - \mathbf{1}_{\{X_2\leq y_2\}}\right)\right] = \Sigma_2(y_1,y_2).$$

Proof. Using the comment before Proposition 4.28, we get:

$$\mathbb{E}\left[\prod_{i\in\{1,2\}}\left(\mathbf{1}_{\{D_i^{(n+1)}\leq d_i\}} - \mathbf{1}_{\{X_i\leq y_i\}}\right)\right] \\ = \mathbb{E}\left[W(X_{[2]})\Psi_{n,-1}(X_{[2]})\right] + \mathbb{E}\left[(1-W(X_{[2]}))\Psi_{n,0}(X_{[2]})\right].$$

We apply Lemma 4.29 twice with $G = W$ and $G = 1 - W$ to get the result. □

4.9 Proof of Theorem 4.23

Recall the definitions of Π_{n+1} and $c_n(y)$ given in (4.61) and (4.67). We define the normalized and centered random process $\hat{\Pi}_{n+1} = (\hat{\Pi}_{n+1}(y) : y \in (0,1))$ by:

$$\hat{\Pi}_{n+1}(y) = \sqrt{n+1}\left[\Pi_{n+1}(y) - c_n(y)\right]. \quad (4.87)$$

Let $U_{n+1} = (U_{n+1}(y) : y \in (0,1))$ be the Hàjek projection of $\hat{\Pi}_{n+1}$:

$$U_{n+1}(y) = \sum_{i=1}^{n+1}\mathbb{E}\left[\hat{\Pi}_{n+1}(y)\middle| X_i\right]. \quad (4.88)$$

Recall Σ defined in Remark 4.24.

Lemma 4.31. *For all $y, z \in (0,1)$, we have:*

$$\lim_{n\to\infty}\mathbb{E}\left[U_{n+1}(y)U_{n+1}(z)\right] = \Sigma(y,z).$$

Proof. Recall (4.71) and (4.72). With $d = D(y)$, we notice that for $y \in (0,1)$:

$$\mathbb{P}\left(D_i^{(n+1)} \leq d\middle| X_j\right) - c_n(y) = \begin{cases} \frac{1}{n}H_n(y,X_j) & \text{if } i \neq j,\\ H_n^\star(y,X_j) & \text{if } i = j.\end{cases}$$

We have:

$$U_{n+1}(y) = (n+1)^{-\frac{1}{2}}\sum_{i=1}^{n+1}\sum_{j=1}^{n+1}\left[\mathbb{P}\left(D_i^{(n+1)}\leq d\middle| X_j\right) - c_n(y)\right] \\ = (n+1)^{-\frac{1}{2}}\sum_{j=1}^{n+1}\left[H_n^\star(y,X_j) + H_n(y,X_j)\right]. \quad (4.89)$$

Let $y, z \in (0,1)$. Since H_n and $H_n^\star$ are centered and $(X_i : i \in \mathbb{N}^*)$ are independent, using (4.89), we obtain that:

$$\begin{aligned}\mathbb{E}\left[U_{n+1}(y)U_{n+1}(z)\right] = \mathbb{E}\left[H_n^\star(y,X_1)H_n^\star(z,X_1)\right] + \mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right] \\ + \mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] + \mathbb{E}\left[H_n^\star(z,X_1)H_n(y,X_1)\right].\end{aligned}$$

Recall $\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_3$ defined in Remark 4.24.

Study of $\mathbb{E}\left[H_n^\star(y,X_1)H_n^\star(z,X_1)\right]$

By Proposition 4.27, see (4.74), and by dominated convergence, we get that:

$$\lim_{n\to\infty} \mathbb{E}\left[H_n^\star(y,X_1)H_n^\star(z,X_1)\right] = \mathbb{E}\left[\left(\mathbf{1}_{\{X_1\leq y\}} - y\right)\left(\mathbf{1}_{\{X_2\leq z\}} - z\right)\right] = \Sigma_1(y,z). \tag{4.90}$$

Study of $\mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right]$

By Proposition 4.27, we have:

$$\begin{aligned}&\mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right] \\ &\quad= \mathbb{E}\left[\left(\frac{D(y)-W(y,X_1)}{D'(y)} + R_n^{4.27}(y,X_1)\right)\left(\frac{D(z)-W(z,X_1)}{D'(z)} + R_n^{4.27}(z,X_1)\right)\right] \\ &\quad= \frac{1}{D'(y)}\frac{1}{D'(z)}\mathbb{E}\left[(D(y)-W(y,X_1))(D(z)-W(z,X_1))\right] + R_n^{(1)} \\ &\quad= \Sigma_2(y_1,y_2) + R_n^{(1)}\end{aligned}$$

where, because of (4.73), $\left|R_n^{(1)}\right| \leq Cn^{-\frac{1}{4}}$ for some finite constant C. We obtain that

$$\lim_{n\to\infty} \mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right] = \Sigma_2(y,z). \tag{4.91}$$

Study of $\mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] + \mathbb{E}\left[H_n^\star(z,X_1)H_n(y,X_1)\right]$

By Proposition 4.27, we have that:

$$\mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] = \mathbb{E}\left[H_n^\star(y,X_1)\frac{1}{D'(z)}(D(z)-W(z,X_1)) + H_n^\star(y,X_1)R_n^{4.27}(z,X_1)\right].$$

Thanks to (4.73) and (4.74), we have $|H_n^\star(y,X_1)| \leq 1$ and $\mathbb{E}\left[\left|R_n^{4.27}(z,X_1)\right|\right] = O\left(n^{-1/4}\right)$. We deduce from Proposition 4.27 and dominated convergence, that:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] &= \mathbb{E}\left[\left(\mathbf{1}_{\{X_1\leq y\}} - y\right)\frac{1}{D'(z)}(D(z)-W(z,X_1))\right] \\ &= \frac{1}{D'(z)}\left(yD(z) - \int_0^y W(z,x)dx\right).\end{aligned}$$

By symmetry, we finally obtain that

$$\lim_{n\to\infty} \mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] + \mathbb{E}\left[H_n^\star(z,X_1)H_n(y,X_1)\right] = \Sigma_3(y,z). \tag{4.92}$$

Conclusion

Combining (4.90), (4.91) and (4.92), we get that

$$\lim_{n\to\infty} \mathbb{E}\left[U_{n+1}(y)U_{n+1}(z)\right] = \Sigma(y,z).$$

□

Lemma 4.32. *We have the following convergence of finite-dimensional distributions:*

$$(U_{n+1}(y) : y \in (0,1)) \xrightarrow[n\to+\infty]{(fdd)} \chi,$$

where $\chi = (\chi(y) : y \in (0,1))$ is a centered Gaussian process with covariance function Σ given in Remark 4.24.

Proof. Let $k \in \mathbb{N}^*$ and $(y_1,\dots,y_k) \in (0,1)^k$. We define the random vector $U^{(k)}_{n+1} = (U_{n+1}(y_i) : i \in [k])$. For all $y \in (0,1)$ and $j \in [n+1]$, we set $g_n(y,X_j) = [H^\star_n(y,X_j) + H_n(y,X_j)]$. Using (4.89), we have:

$$U^{(k)}_{n+1} = (n+1)^{-\frac{1}{2}} \sum_{j=1}^{n+1} Z^{(n+1)}_j,$$

where $Z^{(n+1)}_j = (g_n(y_i,X_j) : i \in [k])$. Notice $(Z^{(n+1)}_j : j \in [n+1])$ is a sequence of independent, uniformly bounded (see Proposition 4.27) and identically distributed random vectors with mean zero and common positive-definite covariance matrix $V_{n+1} = \mathrm{Cov}\left(Z^{(n+1)}_1\right)$. According to Lemma 4.31, we have that $\lim_{n\to\infty} V_{n+1} = \Sigma^{(k)}$, with $\Sigma^{(k)} = (\Sigma(y_i,y_j) : i,j \in [k])$. The multidimensional Lindeberg-Feller condition is trivially satisfied as $(Z^{(n+1)}_j : j \in [n+1])$ are bounded (uniformly in n) with the same distribution. We deduce from the multidimensional central limit theorem for triangular arrays of random variables, see [22] Corollary 18.2, that $(U^{(k)}_{n+1} : n \geq 0)$ converges in distribution towards the Gaussian random vector with distribution $\mathcal{N}(0,\Sigma^{(k)})$. This gives the result. □

Recall $\hat{\Pi}_n(y)$ defined in (4.87). In view of Lemma 4.32 and since $c_n(y) = y + O(1/n)$, in order to prove Theorem 4.23, it is enough to prove that for all $y \in (0,1)$:

$$\hat{\Pi}_{n+1}(y) - U_{n+1}(y) \xrightarrow[n\to+\infty]{L^2} 0. \tag{4.93}$$

Because $\hat{\Pi}_{n+1}$ and U_{n+1} are centered, we deduce from (4.88) that:

$$\mathbb{E}\left[\left(\hat{\Pi}_{n+1}(y) - U_{n+1}(y)\right)^2\right] = \mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] - \mathbb{E}\left[U_{n+1}(y)^2\right].$$

By Lemma 4.31, we have $\mathbb{E}\left[U_{n+1}(y)^2\right] \xrightarrow[n\to\infty]{} \Sigma(y,y)$. So we deduce that the proof of Theorem 4.23 is a complete as soon as the next lemma is proved.

Lemma 4.33. *For all $y \in (0,1)$, we have*

$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] = \Sigma(y,y).$$

Proof. Let $y \in (0,1)$ and $d = D(y)$. We have

$$\begin{aligned}
\mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] &= \frac{1}{n+1} \sum_{i,j=1}^{n+1} \mathbb{E}\left[\left(\mathbf{1}_{\left\{D^{(n+1)}_i \leq d\right\}} - c_n(y)\right)\left(\mathbf{1}_{\left\{D^{(n+1)}_j \leq d\right\}} - c_n(y)\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}_{\left\{D^{(n+1)}_1 \leq d\right\}}\right] - c_n(y)^2 + n\left\{\mathbb{E}\left[\mathbf{1}_{\left\{D^{(n+1)}_1 \leq d\right\}}\mathbf{1}_{\left\{D^{(n+1)}_2 \leq d\right\}}\right] - c_n(y)^2\right\} \\
&= c_n(y) - (n+1)c_n(y)^2 + n\,\mathbb{E}\left[\mathbf{1}_{\left\{D^{(n+1)}_1 \leq d\right\}}\mathbf{1}_{\left\{D^{(n+1)}_2 \leq d\right\}}\right].
\end{aligned}$$

So we get that

$$\mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] = B_n^{(1)} + B_n^{(2)} + B_n^{(3)} + B_n^{(4)},$$

where

$$\begin{aligned}
B_n^{(1)} &= c_n(y) - c_n(y)^2, \\
B_n^{(2)} &= -n(c_n(y) - y)^2, \\
B_n^{(3)} &= n\,\mathbb{E}\left[\left(\mathbf{1}_{\left\{D_1^{(n+1)} \le d\right\}} - \mathbf{1}_{\{X_1 \le y\}}\right)\left(\mathbf{1}_{\left\{D_2^{(n+1)} \le d\right\}} - \mathbf{1}_{\{X_2 \le y\}}\right)\right], \\
B_n^{(4)} &= 2n\,\mathbb{E}\left[\mathbf{1}_{\{X_1 \le y\}}\left(\mathbf{1}_{\left\{D_2^{(n+1)} \le d\right\}} - c_n(y)\right)\right].
\end{aligned}$$

By Equation (4.70), we get $\lim_{n\to\infty} B_n^{(1)} = \Sigma_1(y,y)$ and $\lim_{n\to\infty} B_n^{(2)} = 0$. By Proposition 4.30, we get $\lim_{n\to\infty} B_n^{(3)} = \Sigma_2(y,y)$. Using (4.71), we get $B_n^{(4)} = 2\,\mathbb{E}\left[\mathbf{1}_{\{X_1 \le y\}} H_n(y, X_1)\right]$. By Proposition 4.27 and dominated convergence, we get that:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{X_1 \le y\}} H_n(y, X_1)\right] &= \mathbb{E}\left[\mathbf{1}_{\{X_1 \le y\}}\left(\frac{1}{D'(y)}\left(D(y) - W(y, X_1)\right) + R_n^{4.27}(y, X_1)\right)\right] \\
&\xrightarrow[n\to\infty]{} \frac{1}{D'(y)}\left(yD(y) - \int_0^y W(y,x)dx\right).
\end{aligned}$$

This gives $\lim_{n\to\infty} B_n^{(4)} = \Sigma_3(y,y)$. Then, we get that $\lim_{n\to\infty} \mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] = \Sigma(y,y)$. □

✂ ....

Index of notation

$|A|$ cardinality of set A

$[n] = \{1, \dots, n\}$

$|\beta|$ length of $[n]$-word β

$\mathcal{M}_n$ set of $[n]$-words with all characters distinct

$\mathcal{S}_{n,p} = \{\beta \in \mathcal{M}_n : |\beta| = p\}$

$|\mathcal{S}_{n,p}| = A_n^p = n!/(n-p)!$

$\beta_\ell = \beta_{\ell_1} \dots \beta_{\ell_k}$ for $\ell \in \mathcal{S}_{p,k}$ and $\beta \in \mathcal{S}_{n,p}$

$\mathcal{S}_{n,k}^{\ell,\alpha} = \{\beta \in \mathcal{S}_{n,p} : \beta_\ell = \alpha\}$ for $\alpha \in \mathcal{S}_{n,p}$

$|\mathcal{S}_{n,k}^{\ell,\alpha}| = A_{n-k}^{p-k} = (n-k)!/(n-p)!$

$\mathcal{F}$ set of simple finite graphs

F a simple finite graph (and a finite sequence of simple graphs in Sections 3 to 6 satisfying condition (4.38))

$E(F)$ set of edges of F

$V(F)$ set of vertices of F

$v(F) = |V(F)|$ number of vertices of F

$G_n = G_n(W)$ W-random graph with n vertices associated to the sequence $X = (X_k, k \in \mathbb{N}^*)$ of i.i.d. uniform random variables on $[0,1]$

$t(F,G)$ density of hom. from F to G

$t_{\text{inj}}(F,G)$ density of injective hom.

$t_{\text{ind}}(F,G)$ density of embeddings

$Y^\beta(F,G) = \prod_{\{i,j\}\in E(F)} \mathbf{1}_{\{\{\beta_i,\beta_j\}\in E(G)\}}$

$t_{\text{inj}}(F,G) = |\mathcal{S}_{n,p}|^{-1} \sum_{\beta\in\mathcal{S}_{n,p}} Y^\beta(F,G)$

$\boxed{\ell \in \mathcal{M}_p \text{ and } \alpha \in \mathcal{S}_{p,k} \text{ with } k = |\ell|}$

$t_{\text{inj}}(F^\ell, G^\alpha)$ density of injective hom. such that the labelled vertices ℓ of F, with $V(F) = [p]$, are sent on the labelled vertices α of G, with $V(G) = [n]$

$\mathcal{R}_\ell(F)$ sub-graph of the labeled vertices ℓ of F

$Y^\beta(F^\ell, G^\alpha) = Y^\beta(F, G)$ for $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$

$\hat{Y}^\alpha(F^\ell, G^\alpha) = \prod_{\{i,j\}\in E(\mathcal{R}_\ell(F))} \mathbf{1}_{\{\{\alpha_i,\alpha_j\}\in E(G)\}}$

$Y^\beta(F^\ell, G^\alpha) = \hat{Y}^\alpha(F^\ell, G^\alpha)\, \tilde{Y}^\beta(F^\ell, G^\alpha)$ i.e.:

$$\boxed{Y^\beta = \hat{Y}^\alpha\, \tilde{Y}^\beta}$$

$\tilde{t}_{\text{inj}}(F^\ell, G^\alpha) = |\mathcal{S}_{n,p}^{\ell,\alpha}|^{-1} \sum_{\beta\in\mathcal{S}_{n,p}^{\ell,\alpha}} \tilde{Y}^\beta$

$$\boxed{t_{\text{inj}}(F^\ell, G^\alpha) = \hat{Y}^\alpha\, \tilde{t}_{\text{inj}}(F^\ell, G^\alpha)}$$

$t(F, W)$ hom. densities for graphon W

$t_{\text{ind}}(F, W)$ density of embeddings

$X_\alpha = (X_{\alpha_1}, \ldots, X_{\alpha_k})$ and simil. for X_β

$Z^\beta = Z(X_\beta) = \mathbb{E}[Y^\beta(F^\ell, G_n^\alpha)|X]$

$\tilde{Z}^\beta = \mathbb{E}[\tilde{Y}^\beta(F^\ell, G_n^\alpha)|X]$

$t_x = t_x(F^\ell, W) = \mathbb{E}[Z^\beta|X_\alpha = x]$ and

$t_x = \mathbb{E}[t_{\text{inj}}(F^\ell, G_n^\alpha)|X_\alpha = x]$

$\tilde{t}_x = \tilde{t}_x(F^\ell, W) = \mathbb{E}[\tilde{Z}^\beta|X_\alpha = x]$ and

$\tilde{t}_x = \mathbb{E}[\tilde{t}_{\text{inj}}(F^\ell, G_n^\alpha)|X_\alpha = x]$

$\hat{t}_x = \hat{t}_x(F^\ell, W) = \mathbb{E}[\hat{Y}^\alpha(F^\ell, G_n^\alpha)|X_\alpha = x]$

$\boxed{t_x = \hat{t}_x\, \tilde{t}_x}$ for $x \in [0,1]^k$

$t(F^\ell, W) = \int_{[0,1]^k} t_x\, dx = \mathbb{E}[t_{\text{inj}}(F, G_n)]$

$\hat{t}(F^\ell, W) = \int_{[0,1]^k} \hat{t}_x\, dx = \mathbb{E}[t_{\text{inj}}(\mathcal{R}_\ell(F), G_n)]$ and $\hat{t}(F^\ell, W) = t(\mathcal{R}_\ell(F), W)$

$\Gamma_n^{F,\ell}$ random probability measure:

$$\Gamma_n^{F,\ell}(g) = |\mathcal{S}_{n,k}|^{-1} \sum_{\alpha\in\mathcal{S}_{n,k}} g\left(t_{\text{inj}}(F^\ell, G_n^\alpha)\right)$$

$$\Gamma^{F,\ell}(dx) = \mathbb{E}\left[\hat{\Gamma}_n^{F,\ell}(dx)\right]$$

$\sigma^{F,\ell}(g)^2$ asymptotic variance of

$$\sqrt{n}\left(\Gamma_n^{F,\ell}(g) - \Gamma^{F,\ell}(g)\right)$$

$nD_i^{(n)}$ degree of i in G_n

Π_n empirical CDF of the degrees of G_n

$D(x) = \int_{[0,1]} W(x,y)dy$ degree funct. of W

$\mathcal{H}_{n,d,\delta}(\mathfrak{p}) = \mathbb{P}(X \leq nd + \delta)$ for $X \sim \mathcal{B}(n, \mathfrak{p})$

$\sigma^2_{(x)} = x(1-x)$

$S(x) = \lceil x \rceil - x - \frac{1}{2}$

Φ the CDF of $\mathcal{N}(0,1)$

φ probability distribution density of $\mathcal{N}(01)$

$$\boxed{d = D(y)}$$

$$\boxed{c_n(y) = \mathbb{P}(D_1^{(n+1)} \leq d)}$$

$H_n^\star(y,u) = \mathbb{P}(D_1^{(n+1)} \leq d \,|\, X_1 = u) - c_n(y)$

$\frac{H_n(y,u)}{n} = \mathbb{P}(D_1^{(n+1)} \leq d \,|\, X_2 = u) - c_n(y)$

4.10 Appendix A: Preliminary results for the CDF of binomial distributions

In this section, we study uniform asymptotics for the CDF of binomial distributions. Let $n \in \mathbb{N}^*$, $d \in [0,1]$, $\delta \in \mathbb{R}$ and $\mathfrak{p} \in (0,1)$. We consider the CDF:

$$\mathcal{H}_{n,d,\delta}(\mathfrak{p}) = \mathbb{P}\left(X \leq nd + \delta\right), \tag{4.94}$$

where X a binomial random variable with parameters $(n, \mathfrak{p})$. We denote by Φ the cumulative distribution function of the standard Gaussian distribution and by φ the probability distribution density of the standard Gaussian distribution. We recall (4.68) and (4.69): $\sigma^2_{(x)} = x(1-x)$ for $x \in [0,1]$, and $S(x) = \lceil x \rceil - x - \frac{1}{2}$ for $x \in \mathbb{R}$.

We recall a result from [150], see also [186], Chapter VII: for all $x \in \mathbb{R}$, for all $\mathfrak{p} \in (0,1)$ and $n \in \mathbb{N}^*$ such that $n\sigma^2_{(\mathfrak{p})} \geq 25$, we have:

$$\mathbb{P}\left(X \leq n\mathfrak{p} + \sqrt{n}\,\sigma_{(\mathfrak{p})}x\right) = \Phi(x) + \frac{1}{\sqrt{n}}\mathcal{Q}(\mathfrak{p}, x) + \frac{1}{\sqrt{n}\,\sigma_{(\mathfrak{p})}}S(n\mathfrak{p} + x\sqrt{n}\,\sigma_{(\mathfrak{p})})\varphi(x) + U_n(\mathfrak{p}, x), \tag{4.95}$$

where

$$\mathcal{Q}(\mathfrak{p}, x) = \frac{2\mathfrak{p}-1}{6\sigma_{(\mathfrak{p})}}\varphi''(x) = \frac{2\mathfrak{p}-1}{6\sigma_{(\mathfrak{p})}}(x^2-1)\varphi(x),$$

and

$$|U_n(\mathfrak{p}, x)| \leq \frac{0.2 + 0.3\,|2\mathfrak{p}-1|}{n\sigma^2_{(\mathfrak{p})}} + \exp\left(-\frac{3\sqrt{n}\,\sigma_{(\mathfrak{p})}}{2}\right). \tag{4.96}$$

We use this result to give an approximation of $\mathcal{H}_{n,d,\delta}\left(d + \frac{s}{\sqrt{n}}\right)$.

Proposition 4.34. *Let $\varepsilon_0 \in (0, \frac{1}{2})$ and $K_0 = [\varepsilon_0, 1-\varepsilon_0]$. Let $\alpha > 0$. There exists a positive constant C such that for all $n \geq 2$, $s \in [-\alpha\sqrt{\log(n)}, \alpha\sqrt{\log(n)}]$, $\delta \in [-1,1]$ and $d \in K_0$ such that $d + \frac{s}{\sqrt{n}} \in K_0$, we have:*

$$\mathcal{H}_{n,d,\delta}\left(d + \frac{s}{\sqrt{n}}\right) = \Phi(y_s) + \frac{1}{\sqrt{n}}\frac{\varphi(y_s)}{\sigma_{(d)}}\pi(s,n,d,\delta) + R^{4.34}(s,n,d,\delta),$$

where

$$y_s = \frac{-s}{\sigma_{(d)}} \quad \text{and} \quad \pi(s,n,d,\delta) = \frac{1-2d}{6}(1+2y_s^2) + S(nd+\delta) + \delta \tag{4.97}$$

and

$$\left|R^{4.34}(s,n,d,\delta)\right| \leq C\frac{\log(n)^2}{n}\cdot$$

Proof. In what follows, C denotes a positive constant which depends on ε_0 (but not on $n \geq 2$, $s \in [-\alpha\sqrt{\log(n)}, \alpha\sqrt{\log(n)}]$, $\delta \in [-1,1]$ and $d \in K_0$ such that $d + \frac{s}{\sqrt{n}} \in K_0$) and which may change from line to line. We will also use, without recalling it, that $\sigma_{(.)}$ is uniformly bounded away from 0 on K_0.

For all $\theta \in (0,1]$ such that $d + s\theta \in (0,1)$, we set

$$x_s(\theta) = \frac{-s + \delta\theta}{\sigma_{(d+s\theta)}}\cdot \tag{4.98}$$

Let $\mathfrak{p} = d + \frac{s}{\sqrt{n}}$ and X be a binomial random variable with parameters $(n, \mathfrak{p})$. Because $nd + \delta = n\mathfrak{p} + \sqrt{n}\sigma_{(\mathfrak{p})}\,x_s\left(\frac{1}{\sqrt{n}}\right)$, we can write

$$\mathcal{H}_{n,d,\delta}\left(d + \frac{s}{\sqrt{n}}\right) = \mathbb{P}\left(X \leq n\mathfrak{p} + \sqrt{n}\,\sigma_{(\mathfrak{p})}\,x_s\left(\frac{1}{\sqrt{n}}\right)\right). \tag{4.99}$$

Recall S is defined in (4.69). Using (4.95), we get that:

$$\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right)=\Phi\left(x_s\left(\frac{1}{\sqrt{n}}\right)\right)+\frac{1}{\sqrt{n}}Q^{(1)}_{d,\delta}\left(s,\frac{1}{\sqrt{n}}\right)$$
$$+\frac{1}{\sqrt{n}}Q^{(2)}_{d,\delta}\left(s,\frac{1}{\sqrt{n}}\right)+U_n\left(d+\frac{s}{\sqrt{n}},x_s\left(\frac{1}{\sqrt{n}}\right)\right) \quad (4.100)$$

where for $\theta\in(0,1]$ such that $d+s\theta\in(0,1)$,

$$Q^{(1)}_{d,\delta}(s,\theta)=\frac{2(d+s\theta)-1}{6\sigma_{(d+s\theta)}}\left(x_s(\theta)^2-1\right)\varphi(x_s(\theta)),$$
$$Q^{(2)}_{d,\delta}(s,\theta)=\frac{1}{\sigma_{(d+s\theta)}}S\left(d\theta^{-2}+\delta\right)\varphi(x_s(\theta)).$$

Study of the first term on the right-hand side of (4.100)

Let $\theta\in(0,1/\sqrt{2}]$, and notice that $|\log(\theta)|\geq\log(\sqrt{2})>0$. Recall the definition of $x_s(\theta)$ given by (4.98). By simple computations, we get that for all $0<\theta\leq1/\sqrt{2}$, $|s|\leq\alpha\sqrt{2|\log(\theta)|}$, $|\delta|\leq1$, and $d\in K_0$ such that $d+s\theta\in K_0$,

$$|x_s(\theta)|\leq C\,|\log(\theta)|^{\frac12},\quad |x_s'(\theta)|\leq C\,|\log(\theta)|\quad\text{and}\quad |x_s''(\theta)|\leq C\,|\log(\theta)|^{\frac32}. \quad (4.101)$$

We define the function $\Psi_s(\theta)=\Phi(x_s(\theta))$. Applying Taylor's theorem with the Lagrange form of the remainder for Ψ at $\theta=0$, we have:

$$\Psi_s(\theta)=\Psi_s(0)+\theta\Psi_s'(0)+R^{(1)}_s(\theta),$$

where $R^{(1)}_s(\theta)=\int_0^\theta\Psi_s''(t)(\theta-t)dt$. Recall the definition of $y_s=x_s(0)$ given in (4.97). Elementary calculus gives:

$$\Phi\left(x_s(\theta)\right)=\Phi(x_s(0))+\theta\,x_s'(0)\varphi(x_s(0))+R^{(1)}_s(\theta)$$
$$=\Phi\left(y_s\right)+\theta\left[\frac{(1-2d)}{2\sigma_{(d)}}y_s^2+\frac{\delta}{\sigma_{(d)}}\right]\varphi\left(y_s\right)+R^{(1)}_s(\theta), \quad (4.102)$$

where $R^{(1)}_s(\theta)=\int_0^\theta\left(x_s''(t)-x_s'(t)^2x_s(t)\right)\varphi(x_s(t))\,(\theta-t)\,dt$. Using (4.101) and that $t\varphi(t)$ is bounded, we have:

$$\left|R^{(1)}_s(\theta)\right|\leq C\theta^2(|\log(\theta)|^{\frac32}+|\log(\theta)|^2)\leq C\theta^2|\log(\theta)|^2.$$

Study of the second term on the right-hand side of (4.100)

We have $Q^{(1)}_{d,\delta}(s,\theta)=G_s(\theta)H(x_s(\theta))$ where

$$G_s(\theta)=\frac{2(d+s\theta)-1}{6\sigma_{(d+s\theta)}}\quad\text{and}\quad H(x)=\left(x^2-1\right)\varphi(x).$$

For the first term, we have

$$G_s(\theta)=G_s(0)+R^{(2)}_s(\theta)=\frac{2d-1}{6\sigma_{(d)}}+R^{(2)}_s(\theta),$$

where $R^{(2)}_s(\theta)=\int_0^\theta G_s'(t)dt$. We compute that:

$$G_s'(t)=s\left[\frac{1}{3\sigma_{(d+st)}}+\frac{[2(d+st)-1]^2}{12\sigma^3_{(d+st)}}\right].$$

We obtain that $\left|R_s^{(2)}(\theta)\right| \le C\theta|s| \le C\theta|\log(\theta)|^{\frac{1}{2}}$. For the second term, we have

$$H(x_s(\theta)) = H(x_s(0)) + R_s^{(3)}(\theta) = \left(y_s^2 - 1\right)\varphi\left(y_s\right) + R_s^{(3)}(\theta),$$

where $R_s^{(3)}(\theta) = \int_0^\theta x_s'(t)H'(x_s(t))dt = \int_0^\theta x_s'(t)\left[-x_s(t)^3 + 3x_s(t)\right]\varphi(x_s(t))dt$. Using (4.101) and that $(|t|^3+t)\varphi(t)$ is bounded, we get that $\left|R_s^{(3)}(\theta)\right| \le C\theta|\log(\theta)|$. Finally, we obtain that

$$Q_{d,\delta}^{(1)}(s,\theta) = \frac{2d-1}{6\sigma_{(d)}}\left(y_s^2 - 1\right)\varphi\left(y_s\right) + R_s^{(4)}(\theta) \tag{4.103}$$

with $\left|R_s^{(4)}(\theta)\right| \le C\theta|\log(\theta)|$.

Study of the last term on the right-hand side of (4.100)

We have

$$Q_{d,\delta}^{(2)}(s,\theta) = F_s(\theta)S\left(\frac{d}{\theta^2} + \delta\right)\varphi(x_s(\theta)) \quad \text{with} \quad F_s(\theta) = \frac{1}{\sigma_{(d+s\theta)}}.$$

For the first term on the right-hand side, we have

$$F_s(\theta) = F_s(0) + R_s^{(5)}(\theta) = \frac{1}{\sigma_{(d)}} + R_s^{(5)}(\theta),$$

where $R_s^{(5)}(\theta) = \int_0^\theta F_s'(t)dt = \int_0^\theta \frac{s(2(d+st)-1)}{2\sigma_{(d+st)}^3}dt$. We get that $\left|R_s^{(5)}(\theta)\right| \le C\theta|s| \le C\theta|\log(\theta)|^{\frac{1}{2}}$. For the last term on the right-hand side, we have:

$$\varphi(x_s(\theta)) = \varphi(x_s(0)) + R_s^{(6)}(\theta) = \varphi\left(y_s\right) + R_s^{(6)}(\theta),$$

where $R_s^{(6)}(\theta) = \int_0^\theta x_s'(t)\varphi'(x_s(t))dt = -\int_0^\theta x_s(t)x_s'(t)\varphi(x_s(t))dt$. So, using (4.101) and that $t\varphi(t)$ is bounded, we get that $\left|R_s^{(6)}(\theta)\right| \le C\theta|\log(\theta)|$. Finally, we obtain that

$$Q_{d,\delta}^{(2)}(s,\theta) = \frac{1}{\sigma_{(d)}}S\left(\frac{d}{\theta^2} + \delta\right)\varphi\left(y_s\right) + R_s^{(7)}(\theta), \tag{4.104}$$

where $\left|R_s^{(7)}(\theta)\right| \le C\theta\left|\log(\theta)\right|$, since S is bounded.

Conclusion

We deduce from (4.102), (4.103) and (4.104) that

$$\Phi\left(x_s(\theta)\right) + \theta\, Q_{d,\delta}^{(1)}(s,\theta) + \theta\, Q_{d,\delta}^{(2)}(s,\theta) = \Phi(y_s) + \theta\frac{\varphi(y_s)}{\sigma_{(d)}}\pi\left(s, \frac{1}{\theta^2}, d, \delta\right) + R_s^{(8)}(\theta),$$

where $\left|R_s^{(8)}(\theta)\right| \le C\theta^2|\log(\theta)|^2$. We get the result by taking $\theta = 1/\sqrt{n}$ and using (4.100) and the obvious bound on U_n given by (4.96) so that $|U_n| \le C/n$. □

We state a Lemma which will be useful for the proof of Corollary 4.36.

Lemma 4.35. *Let $y \in [0,1]$ and $\alpha > 0$. For all $n \ge 2$, we have with $d = D(y)$, $A = \alpha\sqrt{\log(n)}$ and $y_s = -s/\sigma_{(d)}$,*

$$\Phi\left(-\frac{A}{\sigma_{(d)}}\right) \le \frac{1}{\alpha\, n^{2\alpha^2}}, \qquad \int_A^{+\infty} s\Phi\left(y_s\right)ds \le \frac{1}{\alpha\, n^{2\alpha^2}} \quad \text{and} \quad \int_A^{+\infty} s^2\varphi(y_s)ds \le \frac{1}{\alpha\, n^{\alpha^2}}.$$

Proof. For all $t \geq 0$, we have

$$\Phi(-t) = \int_t^{+\infty} s\frac{\varphi(s)}{s} ds \leq \frac{1}{t}\int_t^{+\infty} s\varphi(s)ds = \frac{1}{t}\varphi(t). \tag{4.105}$$

Because $\sigma_{(d)} \leq 1/2$, we get with $t = \frac{A}{\sigma_{(d)}}$ the following rough upper bound:

$$\Phi\left(-\frac{A}{\sigma_{(d)}}\right) \leq \frac{\sigma_{(d)}}{A}\varphi\left(\frac{A}{\sigma_{(d)}}\right) \leq \frac{1}{\alpha\, n^{2\alpha^2}}\cdot \tag{4.106}$$

Using again (4.105) and (4.106), we get, for the second inequality that:

$$\int_A^{+\infty} s\Phi\left(y_s\right) ds \leq \sigma_{(d)} \int_A^{+\infty} \varphi(-y_s)ds = \sigma_{(d)}^2 \Phi\left(-\frac{A}{\sigma_{(d)}}\right) \leq \frac{1}{\alpha\, n^{2\alpha^2}}\cdot$$

For the last inequality, we have:

$$\begin{aligned}\int_A^{+\infty} s^2\varphi(y_s)ds = \frac{2\sigma_{(d)}^2}{\sqrt{2\pi}}\int_A^{+\infty}\frac{s^2}{2\sigma_{(d)}^2}\mathrm{e}^{-\frac{s^2}{2\sigma_{(d)}^2}}ds &\leq \frac{4\sigma_{(d)}^2}{\sqrt{2\pi}}\int_A^{+\infty}\mathrm{e}^{-\frac{s^2}{4\sigma_{(d)}^2}}ds\\ &= 4\sqrt{2}\sigma_{(d)}^3\Phi\left(-\frac{A}{\sqrt{2}\sigma_{(d)}}\right)\\ &\leq \frac{1}{\alpha\, n^{\alpha^2}},\end{aligned}$$

where we used $x\mathrm{e}^{-x} \leq 2\mathrm{e}^{-\frac{x}{2}}$ for the first inequality and an inequality similar to (4.106) with $\sigma_{(d)}$ replaced by $\sqrt{2}\sigma_{(d)}$ for the last one. □

For $f \in \mathcal{C}^2([0,1])$, we set $\|f\|_{3,\infty} = \|f\|_\infty + \|f'\|_\infty + \|f''\|_\infty$.

Lemma 4.36. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$ and $\alpha \geq 1$. There exists a positive constant C such that for all $H \in \mathcal{C}^2([0,1])$, $\delta \in [-1,1]$ and $n \geq 2$ such that $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D\big((0,1)\big)$, with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$, we have:*

$$\begin{aligned}\sqrt{n}\int_{-A}^{A} H\left(D^{-1}\left(d+\frac{s}{\sqrt{n}}\right)\right)&\left(\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\leq 0\}}\right)ds\\ &= \frac{H'(y)}{D'(y)}\frac{\sigma_{(d)}^2}{2} + H(y)\left(\frac{1-2d}{2}+\delta+S(nd+\delta)\right) + R_n^{4.36}(H),\end{aligned}$$

where

$$\left|R_n^{4.36}(H)\right| \leq C\,\|H\|_{3,\infty}\, n^{-1/2}\log(n)^3. \tag{4.107}$$

Because of the assumption $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D\big((0,1)\big)$, we need to rule out the cases $y \in \{0,1\}$, so that Lemma 4.36 holds only for $y \in (0,1)$.

Proof. In what follows, C denotes a positive constant which depends on ε_0 and W (but in particular not on $n \geq 2$, $s \in [-\alpha\sqrt{\log(n)}, \alpha\sqrt{\log(n)}]$, $\delta \in [-1,1]$ and $d \in K_0$ such that $d + \frac{s}{\sqrt{n}} \in K_0$) and which may change from lines to lines.

Let $\theta \in (0, 1/\sqrt{2}]$ (we shall take $\theta = 1/\sqrt{n}$ later on) and assume that $|s| \leq \alpha\sqrt{2|\log(\theta)|}$ and $d + s\theta \in K_0$. We set $\Psi(\theta) = H\left(D^{-1}(d+s\theta)\right)$. Notice that $\Psi'(\theta) = \frac{s}{D'\circ D^{-1}(d+s\theta)}H'(D^{-1}(d+s\theta))$. By Taylor's theorem with the Lagrange form of the remainder we have:

$$\Psi(\theta) = \Psi(0) + \theta\Psi'(0) + R_s^{(1)}(\theta) = H(y) + \theta\frac{s}{D'(y)}H'(y) + R_s^{(1)}(\theta) \tag{4.108}$$

where $R_s^{(1)}(\theta) = \int_0^\theta \Psi''(t)(\theta - t)dt$. We have

$$\Psi''(\theta) = s^2 \left[\frac{H''(D^{-1}(d+s\theta))}{(D' \circ D^{-1}(d+s\theta))^2} - \frac{H'(D^{-1}(d+s\theta))(D'' \circ D^{-1}(d+s\theta))}{(D' \circ D^{-1}(d+s\theta))^3} \right].$$

Thus, we get that $\left| R_s^{(1)}(\theta) \right| \leq C\ (\|H'\|_\infty + \|H''\|_\infty)\ s^2\theta^2 \leq C\,\|H\|_{3,\infty}\,\theta^2 |\log(\theta)|$. Choosing $\theta = 1/\sqrt{n}$, we deduce from (4.108) that:

$$H\left(D^{-1}\left(d + \frac{s}{\sqrt{n}}\right)\right) = H(y) + \frac{1}{\sqrt{n}}\frac{s}{D'(y)}H'(y) + R_s^{(1)}\left(\frac{1}{\sqrt{n}}\right), \tag{4.109}$$

where $\left| R_s^{(1)}\left(1/\sqrt{n}\right) \right| \leq C\,\|H\|_{3,\infty}\, \log(n)/n$. Recall the definition of y_s and $\pi(s,n,d,\delta)$ given by (4.97). By Proposition 4.34 and equation (4.109), we get that:

$$\begin{aligned}
&\sqrt{n}\,H\left(D^{-1}\left(d+\frac{s}{\sqrt{n}}\right)\right)\left(\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\leq 0\}}\right)\\
&= \sqrt{n}\left(H(y) + \frac{1}{\sqrt{n}}\frac{s}{D'(y)}H'(y)\right)\left(\left(\Phi(y_s) - \mathbf{1}_{\{s\leq 0\}}\right) + \frac{1}{\sqrt{n}}\frac{\varphi(y_s)}{\sigma_{(d)}}\pi(s,n,d,\delta)\right) + R_n^{(0)}(s)\\
&= \sqrt{n}\,H(y)\Delta^{(1)}(s) + \frac{H'(y)}{D'(y)}\Delta^{(2)}(s) + \frac{H(y)}{\sigma_{(d)}}\Delta^{(3)}(s) + R_n^{(0)}(s) + \hat{R}_n^{(0)}(s),
\end{aligned} \tag{4.110}$$

where

$$\Delta^{(1)}(s) = \left(\Phi(y_s) - \mathbf{1}_{\{s\leq 0\}}\right), \quad \Delta^{(2)}(s) = s\left(\Phi(y_s) - \mathbf{1}_{\{s\leq 0\}}\right), \quad \Delta^{(3)}(s) = \varphi(y_s)\pi(s,n,d,\delta),$$

$$\begin{aligned}
\left|R_n^{(0)}(s)\right| &\leq \sqrt{n}\|H\|_\infty |R^{4.34}(s,n,d,\delta)| + \sqrt{n}|R_s^{(1)}\left(1/\sqrt{n}\right)| + \sqrt{n}|R^{4.34}(s,n,d,\delta)R_s^{(1)}\left(1/\sqrt{n}\right)|\\
&\leq C\,\|H\|_{3,\infty}\frac{\log(n)^2}{\sqrt{n}}
\end{aligned} \tag{4.111}$$

and

$$\left|\hat{R}_n^{(0)}(s)\right| = \left|\frac{1}{\sqrt{n}}\frac{H'(y)}{\sigma_{(d)}D'(y)}s\varphi(y_s)\pi(s,n,d,\delta)\right| \leq C\,\|H\|_{3,\infty}\frac{\sqrt{\log(n)}}{\sqrt{n}}. \tag{4.112}$$

Study of $\int_{-A}^{A} \Delta^{(1)}(s)ds$

Since $\Delta^{(1)}$ is an odd integrable function on $\mathbb{R}^*$, we get that:

$$\int_{-A}^{A} \Delta^{(1)}(s)ds = 0. \tag{4.113}$$

Study of $\int_{-A}^{A} \Delta^{(2)}(s)ds$

Because $\Delta^{(2)}$ is integrable and $\int_\mathbb{R} \Delta^{(2)}(s)ds = \sigma_{(d)}^2/2$, we get that

$$\int_{-A}^{A} \Delta^{(2)}(s)\,ds = \frac{\sigma_{(d)}^2}{2} + R_n^{(2)}, \quad \text{with} \quad R_n^{(2)} = -2\int_A^{+\infty} s\Phi(y_s)ds. \tag{4.114}$$

Using Lemma 4.35, we get that

$$|R_n^{(2)}| \leq C\,n^{-2\alpha^2}. \tag{4.115}$$

Study of $\int_{-A}^{A} \Delta^{(3)}(s)ds$

We have, using (4.97), that:

$$\Delta^{(3)}(s) = \varphi(y_s)\left(\frac{1-2d}{6}(1+2y_s^2) + \delta + S(nd+\delta)\right).$$

By elementary calculus, we have that:

$$\int_{\mathbb{R}} \varphi(y_s)ds = \int_{\mathbb{R}} y_s^2\varphi(y_s)ds = \sigma_{(d)}.$$

We get that:

$$\int_{-A}^{A} \Delta^{(3)}(s)ds = \sigma_{(d)}\left[\frac{1-2d}{2} + \delta + S\left(nd+\delta\right)\right] + R_n^{(3)}, \tag{4.116}$$

where

$$R_n^{(3)} = -2\sigma_{(d)}\left(\frac{1-2d}{6} + \delta + S(nd+\delta)\right)\Phi\left(-\frac{A}{\sigma_{(d)}}\right) - 2\,\frac{1-2d}{3}\int_{A}^{+\infty} y_s^2\varphi(y_s)ds. \tag{4.117}$$

Using Lemma 4.35 and since $|2d-1| \leq 1$, $|\delta| \leq 1$ and S is bounded by 1, we have that:

$$|R_n^{(3)}| \leq Cn^{-\alpha^2}. \tag{4.118}$$

Conclusion

Using (4.110), (4.113), (4.114), (4.116), we deduce that

$$\sqrt{n}\int_{-A}^{A} H\left(D^{-1}\left(d+\frac{s}{\sqrt{n}}\right)\right)\left(\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\leq 0\}}\right)ds$$
$$= \frac{H'(y)}{D'(y)}\frac{\sigma_{(d)}^2}{2} + H(y)\left[\frac{1-2d}{2} + \delta + S\left(nd+\delta\right)\right] + R_n^{4.36}(H),$$

where $R_n^{4.36}(H) = \int_{-A}^{A}(R_n^{(0)}(s) + \hat{R}_n^{(0)}(s))\,ds + (H'(y)/D'(y))R_n^{(2)} + (H(y)/\sigma_{(d)})R_n^{(3)}$. Using the upper bounds (4.111) and (4.112) (to be integrated over $[-A,A]$), (4.115) and (4.118) with $\alpha \geq 1$, we get that $|R_n^{4.36}(H)| \leq C\,\|H\|_{3,\infty}\log(n)^3/\sqrt{n}$. □

We give a direct application of the previous lemma.

Lemma 4.37. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$ and $\alpha \geq 1$. There exists a positive constant C such that for all $G \in \mathcal{C}^2([0,1])$, $\delta \in [-1,1]$, $n \geq 2$ such that $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D((0,1))$, with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$, we have:*

$$n\int_{0}^{1} G(x)\left(\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x\leq y\}}\right)\mathbf{1}_{\{\sqrt{n}|D(x)-d|\leq A\}}dx$$
$$= \frac{G'(y)D'(y) - G(y)D''(y)}{D'(y)^3}\frac{\sigma_{(d)}^2}{2} + \frac{G(y)}{D'(y)}\left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.37}(G),$$

where

$$\left|R_n^{4.37}(G)\right| \leq C\,\|G\|_{3,\infty}\,n^{-1/2}\log(n)^3.$$

Proof. Let G be a function in $\mathcal{C}^2([0,1])$. Define the function H on $[0,1]$ by $H(z) = \frac{G(z)}{D'(z)}$ for all $z \in [0,1]$. Use the change of variables $s = \sqrt{n}\,(D(x) - d)$ to get that:

$$\int_0^1 G(x)\left(\mathcal{H}_{n,d,\delta}(D(x)) - \mathbf{1}_{\{x\leq y\}}\right)\mathbf{1}_{\{\sqrt{n}|D(x)-d|\leq A\}}dx$$
$$= \frac{1}{\sqrt{n}}\int_{-A}^{A} H\left(D^{-1}\left(d+\frac{s}{\sqrt{n}}\right)\right)\left(\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\leq 0\}}\right)ds.$$

By Lemma 4.36, we obtain that:

$$n\int_0^1 G(x)\left(\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x\leq y\}}\right)\mathbf{1}_{\left\{\sqrt{n}|D(x)-d|\leq A\right\}}dx$$
$$= \frac{H'(y)}{D'(y)}\frac{\sigma^2_{(d)}}{2} + H(y)\left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.36}(H)$$
$$= \frac{G'(y)D'(y) - G(y)D''(y)}{D'(y)^3}\frac{\sigma^2_{(d)}}{2} + \frac{G(y)}{D'(y)}\left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.36}(G/D').$$

Set $R_n^{4.37}(G) = R_n^{4.36}\,(G/D')$ and use (4.107) to conclude. □

Lemma 4.38. *Let $y \in (0,1)$ and $\alpha > 0$. For all $u \in (0,1)$, $\delta \in [-1,1]$ and $n \in \mathbb{N}^*$ such that $\sqrt{n}|u-d| \geq A$ with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$, we have*

$$\left|\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\leq d\}}\right| \leq n^{-\alpha+2}.$$

Proof. Let X be a binomial random variable with parameters (n,u). Assume first that $u \geq d + \frac{A}{\sqrt{n}}$. Let $\lambda \geq 0$. Using Chernov's inequality, we get:

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\leq d\}} = \mathbb{P}\left(X \leq nd+\delta\right) \leq \mathrm{e}^{\lambda(nd+\delta)}\,\mathbb{E}\left[\mathrm{e}^{-\lambda X}\right] = \exp\left[\lambda(nd+\delta) + n\Psi(\lambda)\right], \tag{4.119}$$

with $\Psi(\lambda) = \log(1+u(\mathrm{e}^{-\lambda}-1))$. By Taylor's theorem with the Lagrange form of the remainder, we have

$$\Psi(\lambda) = \Psi(0) + \lambda\Psi'(0) + R(\lambda) = 0 - u\lambda + R(\lambda), \tag{4.120}$$

where $R(\lambda) = \int_0^\lambda(\lambda-t)\Psi''(t)dt$. Because $\Psi''(t) \geq 0$ and $\Psi''(t) = \frac{(1-u)u\mathrm{e}^{-\lambda}}{\left(1+u(\mathrm{e}^{-\lambda}-1)\right)^2} \leq \frac{1}{4}$ (applying the following inequality $\frac{xy}{(x+y)^2} \leq \frac{1}{4}$ with $x = 1-u$ and $y = u\mathrm{e}^{-\lambda}$), we get that $|R(\lambda)| \leq \frac{\lambda^2}{8} \leq \lambda^2$. Finally, applying (4.120) with $\lambda = \sqrt{\frac{\log(n)}{n}}$, we get that

$$n\Psi(\lambda) = -u\sqrt{n\log(n)} + R^{(2)}(n), \tag{4.121}$$

with $|R^{(2)}(n)| \leq \log(n)$. Using (4.119) and (4.121), we get that

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\leq d\}} \leq \exp\left[\sqrt{\frac{\log(n)}{n}}(nd+\delta) - u\sqrt{n\log(n)} + R^{(2)}(n)\right]$$
$$= \exp\left[\sqrt{n\log(n)}(d-u) + R^{(3)}(n)\right],$$

where $|R^{(3)}(n)| \leq 2\log(n)$, since $|\delta| \leq 1$. Because $d - u \leq \frac{-A}{\sqrt{n}}$ with $A = \alpha\sqrt{\log(n)}$, we have that

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\leq d\}} \leq \mathrm{e}^{-\alpha\log(n)+R^{(3)}(n)} \leq \mathrm{e}^{(-\alpha+2)\log(n)} = n^{-\alpha+2}.$$

In the case where $u \leq d - \frac{A}{\sqrt{n}}$, we have that

$$\begin{aligned} 0 \geq \mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u \leq d\}} &= \mathbb{P}\left(X \leq nd + \delta\right) - 1 \\ &\geq -\mathbb{P}\left(X \geq nd + \delta\right) \\ &= -\mathbb{P}\left(n - X \leq n(1-d) - \delta\right). \end{aligned}$$

Since $n - X$ is a binomial random variable with parameters $(n, 1-u)$, using similar argument as in the first part of the proof (with u and X replaced by $1-u$ and $n-X$), we get that, for $u \leq d - \frac{A}{\sqrt{n}}$:

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u \leq d\}} \geq -n^{-\alpha+2}.$$

We deduce that $\left|\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u \leq d\}}\right| \leq n^{-\alpha+2}$. □

The following lemma is a direct application of Lemma 4.38 with $u = D(x)$.

Lemma 4.39. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$ and $\alpha \geq 1$. For all $G \in \mathcal{B}([0,1])$, $\delta \in [-1,1]$ and $n \in \mathbb{N}^*$, we have with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$:*

$$n \int_0^1 G(x) \left|\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x \leq y\}}\right| \mathbf{1}_{\{\sqrt{n}|D(x)-d| \geq A\}} dx = R_n^{4.39}(G),$$

where

$$\left|R_n^{4.39}(G)\right| \leq \|G\|_\infty \, n^{-\alpha+3}.$$

Combining Lemma 4.37 with Lemma 4.39 for $\alpha = 3$, we deduce the following proposition.

Proposition 4.40. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$. There exists a positive constant C such that for all $G \in \mathcal{C}^2([0,1])$, $\delta \in [-1,1]$ and $n \in \mathbb{N}^*$ such that $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D((0,1))$, with $d = D(y)$ and $A = 4\sqrt{\log(n)}$, we have:*

$$n \int_0^1 G(x) \left(\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x \leq y\}}\right) dx$$
$$= \frac{G'(y)D'(y) - G(y)D''(y)}{D'(y)^3} \frac{\sigma^2_{(d)}}{2} + \frac{G(y)}{D'(y)} \left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.40}(G),$$

with

$$\left|R_n^{4.40}(G)\right| \leq C \|G\|_{3,\infty} \, n^{-\frac{1}{4}}. \tag{4.122}$$

4.11 Appendix B: Proof of Proposition 4.28

We first state a preliminary lemma in Section 4.11.1 and then provide the proof of Proposition 4.28 in Section 4.11.2.

4.11.1 A preliminary result

For $\mathrm{y} = (y_1, y_2) \in [0,1]^2$, let $M(\mathrm{y})$ be the covariance matrix of a pair (Y_1, Y_2) of Bernoulli random variables such that $\mathbb{P}(Y_i = 1) = D(y_i)$ for $i \in \{1,2\}$ and $\mathbb{P}(Y_1 = Y_2 = 1) = \int_{[0,1]} W(y_1, z) W(y_2, z) \, dz$.

Lemma 4.41. *Assume that W satisfies condition (4.62). There exists $\varepsilon' > 0$ such that for all $\mathrm{y} \in [0,1]^2$, we have $\det(M(\mathrm{y})) > \varepsilon'$.*

Proof. Let $\mathbb{M}_2$ be the set of matrices of size 2×2, and $\| \cdot \|_\infty$ be the norm on $\mathbb{M}_2$ defined in (4.127). We consider the closed set on $\mathbb{M}_2$:

$$\mathcal{F} = \mathcal{F}_+ \bigcup \mathcal{F}_- \quad \text{where} \quad \mathcal{F}_\pm = \left\{ r(I_2 \pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}); \, r \in [0, 1/4] \right\}$$

where $I_2 \in \mathbb{M}_2$ is the identity matrix. Notice $\mathcal{F}$ is the set of all covariance matrices of pairs of Bernoulli random variables having determinant equal to 0. Since the determinant is a continuous real-valued function on $\mathbb{M}_2$, to prove Lemma 4.41, it is enough to prove that for all $\mathrm{y} = (y, y') \in [0,1]^2$ and all $M_0 \in \mathcal{F}$:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4. \tag{4.123}$$

We set $p = D(y)$, $p' = D(y')$ and $\alpha = \int_{[0,1]} W(y, z) W(y', z) \, dz$ so that:

$$M(\mathrm{y}) = \begin{pmatrix} p(1-p) & \alpha - pp' \\ \alpha - pp' & p'(1-p') \end{pmatrix}.$$

The proof of (4.123) is divided into three cases. Recall that W satisfies condition (4.62). Without loss of generality, we can assume that $p \leq p'$ and thus:

$$\varepsilon \leq p \leq p' \leq 1 - \varepsilon. \tag{4.124}$$

Since $(1 - W(y, z))(1 - W(y', z))$ is non-negative, by integrating with respect to z over $[0, 1]$, we get that $\alpha \geq p + p' - 1$. Using that $W \leq 1 - \varepsilon$, we deduce, denoting by $x^+ = \max(x, 0)$ the positive part of $x \in \mathbb{R}$, that:

$$(p + p' - 1)_+ \leq \alpha \leq (1 - \varepsilon)p. \tag{4.125}$$

The case $M_0 \in \mathcal{F}_+$

Recall that $p \leq p'$. If $|r - p(1-p)| \geq \varepsilon^2/4$, then, by considering the first term on the diagonal, we have $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

If $|r - p(1-p)| \leq \varepsilon^2/4$, then, by considering the off-diagonal term, we have:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq |\alpha - pp' - r|.$$

For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, we get, using that $\alpha \leq (1-\varepsilon)p$ and $p \leq p'$:

$$\begin{aligned} \alpha - pp' - r &\leq (1-\varepsilon)p - p^2 - p(1-p) - \delta' \\ &\leq -\varepsilon^2 + \varepsilon^2/4 = -3\varepsilon^2/4. \end{aligned}$$

We deduce that (4.123) holds if $M_0 \in \mathcal{F}_+$.

The case $|1 - p - p'| > \varepsilon/2$ and $M_0 \in \mathcal{F}_-$

If $|r - p(1-p)| \geq \varepsilon^2/4$, then, by considering the first term on the diagonal, we have $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

If $|r - p(1-p)| \leq \varepsilon^2/4$, then, by considering the off-diagonal term, we have:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq |\alpha - pp' + r|.$$

Assume first that $1 - p - p' > \varepsilon/2$. For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, we get, using $\alpha \geq 0$, that:

$$\begin{aligned} \alpha - pp' + r &\geq p(1 - p - p') + \delta' \\ &\geq \varepsilon^2/2 - \varepsilon^2/4 = \varepsilon^2/4. \end{aligned}$$

Assume then that $1-p-p' < -\varepsilon/2$. For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, we get, using the lower bound $\alpha \geq p + p' - 1$ from (4.125), that:

$$\begin{aligned} \alpha - pp' + r &\geq (1-p)(p+p'-1) + \delta' \\ &\geq \varepsilon^2/2 - \varepsilon^2/4 = \varepsilon^2/4. \end{aligned}$$

We get $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

We deduce that (4.123) holds if $|1-p-p'| > \varepsilon/2$ and $M_0 \in \mathcal{F}_-$.

The case $|1-p-p'| \leq \varepsilon/2$ and $M_0 \in \mathcal{F}_-$

Applying Lemma 4.42 below, with $f = W(y, \cdot)$, $g = W(y', \cdot)$ and $\delta = 1-p-p'$, we get that:

$$\alpha \geq (1-\varepsilon)(\varepsilon - \delta). \tag{4.126}$$

If $|r - p(1-p)| \geq \varepsilon^2/4$, then, by considering the first term on the diagonal, we have $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

If $|r - p(1-p)| \leq \varepsilon^2/4$, then, by considering the off-diagonal term, we have:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq |\alpha - pp' + r|.$$

For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, using (4.126), we get that:

$$\begin{aligned} \alpha - pp' + r &= \alpha - p(1-p-\delta) + p(1-p) + \delta' \\ &\geq (1-\varepsilon)\varepsilon - \delta(1-\varepsilon-p) + \delta' \\ &\geq (1-\varepsilon)\varepsilon - (1-2\varepsilon)\varepsilon/2 - \varepsilon^2/4 \geq \varepsilon^2/4. \end{aligned}$$

We deduce that (4.123) holds if $|1-p-p'| \leq \varepsilon/2$ and $M_0 \in \mathcal{F}_-$.

Conclusion

Since (4.123) holds when $M_0 \in \mathcal{F}_+$, when $M_0 \in \mathcal{F}_-$ and either $|1-p-p'| > \varepsilon/2$ or $|1-p-p'| \leq \varepsilon/2$, we deduce that (4.123) holds under the condition of Lemma 4.41. □

Lemma 4.42. *Let $\varepsilon \in (0, 1/2)$, $\delta \in [-\varepsilon/2, \varepsilon/2]$, $f, g \in \mathcal{B}([0,1])$ such that $0 \leq f, g \leq 1-\varepsilon$ and $\int_{[0,1]}(f+g) = 1-\delta$. Then we have $\int_{[0,1]} fg \geq (1-\varepsilon)(\varepsilon-\delta)$, and this lower bound is sharp.*

Proof. Set $f_1 = \min(f,g)$ and $g_1 = \max(f,g)$ so that $0 \leq f_1 \leq g_1 \leq 1-\varepsilon$ and $\int_{[0,1]}(f_1+g_1) = 1-\delta$ and $\int_{[0,1]} f_1\, g_1 = \int_{[0,1]} fg$. Set $h = \min(f_1, (1-\varepsilon-g_1))$ as well as $f_2 = f_1 - h$ and $g_2 = g_1 + h$ so that $0 \leq f_2 \leq g_2 \leq 1-\varepsilon$, $\int_{[0,1]}(f_2+g_2) = 1-\delta$ and

$$\int_{[0,1]} f_2\, g_2 = \int_{[0,1]} (f_1 - h)\,(g_1 + h) = \int_{[0,1]} f_1\, g_1 - \int_{[0,1]} (h(g_1 - f_1) + h^2) \leq \int_{[0,1]} f_1\, g_1 = \int_{[0,1]} fg.$$

Since by construction either $f_2(x) = 0$ or $g_2(x) = 1-\varepsilon$, we deduce that:

$$\int_{[0,1]} fg \geq \int_{[0,1]} f_2\, g_2 \geq (1-\varepsilon) \int_{[0,1]} f_2 = (1-\varepsilon)\left(1 - \delta - \int_{[0,1]} g_2\right) \geq (1-\varepsilon)(\varepsilon - \delta).$$

To see this lower bound is sharp, consider $g = 1-\varepsilon$ and $f = \varepsilon - \delta$. □

4.11.2 Proof of Proposition 4.28

We set:

$$\hat{Z}_n = (n-1)^{-1/2} M(\mathrm{x})^{-1/2} (\mathcal{D}^{(n+1)} - \mu(\mathrm{x})),$$

which is, conditionally on $\{X_{[2]} = \mathrm{x}\}$, distributed as the normalized and centered sum of $n-1$ independent random variables distributed as $Y = (Y_1, Y_2)$, with Y_1 and Y_2 Bernoulli random variables such that $\mathbb{E}[Y] = \mu(\mathrm{x})/(n-1)$ and $\mathrm{Cov}(Y,Y) = M(\mathrm{x})$.

Using Theorem 3.5 from [48] or Theorem 1.1 from [20], we get that:

$$\sup_{K \in \mathcal{K}} \left| \mathbb{P}\left(\hat{Z}_n \in K \middle| X_{[2]} = \mathrm{x} \right) - \mathbb{P}(Z \in K) \right| \leq 115\sqrt{2}\,\gamma,$$

where

$$\gamma = (n-1)\mathbb{E}\left[\left| (n-1)^{-1/2} M(\mathrm{x})^{-1/2} (Y - \mathbb{E}[Y]) \right|^3 \right].$$

Let $\|\cdot\|_\infty$ denote the matrix norm on the set $\mathbb{M}_2$ of real matrices of dimension 2×2 induced by the maximum vector norm on $\mathbb{R}^2$, which is the maximum absolute row sum:

$$\| M \|_\infty = \max_{1 \leq i \leq 2} \sum_{j=1}^{2} |M(i,j)|, \quad \text{for all } M \in \mathbb{M}_2. \tag{4.127}$$

Recall that $\|\cdot\|_\infty$ is an induced norm (that is $\| AB \|_\infty \leq \| A \|_\infty \| B \|_\infty$). For $M \in \mathbb{M}_2$ and $\mathrm{x} \in \mathbb{R}^2$, we have $|M\mathrm{x}| \leq \sqrt{2}\, \| M \|_\infty |\mathrm{x}|$. If $M \in \mathbb{M}_2$ is symmetric positive definite (which is only used for the second inequality and the equality), we get:

$$\| M \|_\infty^{1/2} \leq \| M^{1/2} \|_\infty \leq \sqrt{2} \| M \|_\infty^{1/2} \quad \text{and} \quad \| M^{-1} \|_\infty = \frac{\| M \|_\infty}{|\det(M)|}. \tag{4.128}$$

We deduce that if $M \in \mathbb{M}_2$ is symmetric positive definite, then

$$\| M^{-1/2} \|_\infty \leq \sqrt{2} |\det(M)|^{-1/2} \| M \|_\infty^{1/2}.$$

We obtain that for $n \geq 2$:

$$\begin{aligned}
\gamma &\leq 2^3 (n-1)^{-1/2} \| M(\mathrm{x}) \|_\infty^{3/2} \det(M(\mathrm{x}))^{-3/2}\, \mathbb{E}\left[|Y - \mathbb{E}[Y]|^3 \right] \\
&\leq 2^{5/2} n^{-1/2} \det(M(\mathrm{x}))^{-3/2}\, \mathbb{E}\left[|Y_1 - \mathbb{E}[Y_1]|^3 + |Y_2 - \mathbb{E}[Y_2]|^3 \right] \\
&\leq 2^{3/2} n^{-1/2} \det(M(\mathrm{x}))^{-3/2},
\end{aligned}$$

where we used that $\| M(\mathrm{x}) \|_\infty \leq 1/2$ for the second inequality and the convex inequality $(x+y)^p \leq 2^{p-1}(x^p + y^p)$ for the third and that $|Y_i - \mathbb{E}[Y_i]| \leq 1$ so that $\mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^3] \leq \mathrm{Var}\,(Y_i) \leq 1/4$. We deduce from Lemma 4.41 that there exists $C_0 > 0$ such that for all $\mathrm{x} = (x_1, x_2) \in [0,1]^2$ with $x_1 \neq x_2$ and all $n \geq 2$:

$$\sup_{K \in \mathcal{K}} \left| \mathbb{P}\left(\hat{Z}_n \in K \middle| X_{[2]} = \mathrm{x} \right) - \mathbb{P}(Z \in K) \right| \leq C_0\, n^{-1/2}.$$

To conclude, replace the convex set K in this formula by the convex set $\frac{M(\mathrm{x})^{-\frac{1}{2}}}{\sqrt{n-1}}(K - \mu(\mathrm{x}))$.