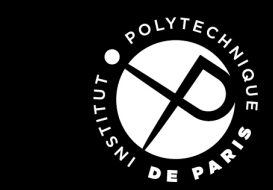
Graphon estimation in bipartite networks

Etienne Donier-Meroz



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAG010

Thèse de doctorat

Graphon estimation in bipartite networks

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique (ENSAE)

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 23 novembre 2023, par

ETIENNE DONIER-MEROZ

Composition du Jury :

Alexandre Tsybakov Professeur, CREST, ENSAE, IP Paris	Président
Alain Célisse Professeur, Université Paris 1	Rapporteur et examinateur
Ery Arias-Castro Professeur, University of California, San Diego	Rapporteur
Catherine Matias Directrice de recherche, CNRS	Examinatrice
Christophe Denis Maître de conférence, LPSM Sorbonne-Université	Examinateur
Arnak Dalalyan Professeur, CREST, ENSAE, IP Paris	Directeur de thèse
Francis Kramarz Professeur, CREST, ENSAE, IP Paris	Co-directeur de thèse

Remerciements

Ma première pensée va sans aucun doute à Dieu, car c'est seulement grâce à lui que j'ai la vie et que suis parvenu jusqu'à ce niveau d'étude, et c'est pour moi une joie de lui rendre gloire en lui consacrant les quelques premiers mots de mon manuscrit de thèse qui reflète l'accomplissement de mes études de mathématiques, dans lesquelles il a été plus que présent.

Je remercie tous les membres du Jury, d'avoir participé à ma soutenance de thèse, merci pour votre temps et votre intérêt pour nos travaux. Je remercie particulièrment Alain et Ery d'avoir accepté d'être mes rapporteurs pour cette thèse, merci pour vos commentaires et remarques utiles et pour tout le temps que vous avez consacré à lire mon manuscrit.

Merci à tous mes collègues du CREST pour la bonne ambiance humaine qui n'est possible que par vous. Vous savez tous allier travail et détente, respect et communion. Je remercie particulièrement Victor-Emmanuel, pour le suivi de mon doctorat et ta disponibilité; merci à Nicolas pour tous tes conseils en programmation, pour ta gentillesse et toutes les parties de ping-pong. Merci aussi à Guillaume, Cristina, Lucas, Jules pour avoir participé au bon déroulement de ma thèse et de mon enseignement. Merci à la formidable équipe administrative qui m'a accompagné durant ma thèse, merci aussi Stephane pour la direction de l'EDMH, pour tes conseils et ta bienveillance.

Merci à tous les post-doc, doctorant et anciens doctorants du CREST, pour toutes les discussions et les temps qu'on a pu passer ensemble, en particulier, merci Julien (pour ta colaboration en TD), Flore, Suzanne, Younes, Théo, Meyer, Gabriel, Nayel, Lu, Jordan et Arshak (qui est aussi un excellent co-équipier de foot). Un grand merci aux membres et anciens membres du bureau 3012 pour l'ambiance extraordinare et le soutien que vous avez été pour moi durant ma thèse. Je remercie notamment Avetik, Amir mais aussi Hafedh, Nina, et surtout Clara et Hugo pour tous ces bons moments passés ici, vous avez été d'un soutien extrêmement important pour moi.

Un grand merci à ma famille, pour le soutien que vous êtes dans ma vie de tous les jours. Merci à mes parents, Chantal et Daniel, c'est aussi grâce à vous que j'écris ces lignes, ma petite soeur Julie, ma grande soeur Brigitte, mon beau frère Tristan, et mon petit neveu nouveau né Lucas pour la joie que tu as apporté dans notre maison.

Un énorme Merci pour tous mes amis et les membres de mon église, merci pour votre

soutien en prière qui a fait la différence durant ces trois années et plus encore.

Merci à Xavier, Phillipe et Francis pour votre investissement profond dans ma thèse, qui n'aurait pas pu être possible sans vous. Merci pour votre aide, vos conseils, vos relectures, votre colaboration et tout ce temps investi dans nos travaux de recherche.

Le dernier, mais pas des moindres... Je te remercie Arnak, je n'ai pas vraiment les mots pour te dire à quel point je suis reconnaissant pour ta direction durant cette thèse, alors je dirai avec sincérité que tu es sans aucun doute le meilleur directeur de thèse que j'aurais pu espérer avoir. Tes qualités en tant que mathématicien et en tant que personne sont remarquables. J'ai vraiment l'impression d'avoir progressé dans ce milieu inconnu qu'était la recherche grâce à toi. Merci pour ton soutien constant, tu as su parfaitement me guider, me rassurer et me motiver quand il le fallait. Et en plus de cela, j'ai eu ce grand privilège d'être ton co-équipier de foot, ce qui reste pour moi une chose que je n'aurais pas même imaginée.

Contents

Chapter 1

Introduction

1.1 Modeling data by Networks and graphs

In statistics, more and more datasets can be represented as a form of a graph (also called network), where individuals make links with each other, that we often call interactions. Famous research topics are for example social networks studies, international trade exchange, biology, consumption market and so on. These structures are valuable because they are easy to visualize. This section is devoted to a reminder of some basic notions and formalizations about network datasets, and some classical models similar to those studied in the present study.

1.1.1 Graphs

Mathematically speaking, a graph $\mathcal{G} = (V, E)$ is the data of a set of points V, called nodes or vertices, and a set E of pairs of these vertices, called the edges. We often assume that $V = \{1, \dots, n\}$ where n is the number of vertices. The edges represents connections or interactions between the entities. For example, $\{1, 2\} \in E$ means that there is an edge between vertex 1 and 2. We assume that edges are not directed, meaning that if i is connected to j, then j is connected to i. We also prohibit a node to be linked with itself. A graph can be fully understand thank to a mathematical object called adjacency matrix. The adjacency matrix $\mathbf{A}$ of a graph $\mathcal{G} = (V, E)$ is a square matrix of size $n \times n$ such that

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

More generally, some graphs could have labeled edges, then the adjacency matrix $\mathbf{A}$ is no longer composed of 0 or 1 but of real entries. In the context of undirected graphs, adjacency matrices are always symmetric and has null diagonal.

In a study conducted by [OW14b], the Political Weblog data was analyzed to examine the level of interaction between liberal and conservative blogs during the 2004 US presidential election. In this context, the weblogs were represented as vertices in a network, and an edge was established between two weblogs if either one of them had a link to the other on their front page.

Additionally, in order to understand how covariates can affect the structure of a network, the authors analyzed a student friendship network from the US National Longitudinal Study of Adolescent Health (Add Health). In this study, students were asked to provide information about their gender, race, and school year (grades 7–12), and to nominate up to 5 friends of each gender. The vertices of resulting network are the student, with a link present whenever either of a pair of students nominated the other as a friend.

When discussing graphs, another matrix of interest that contains important structural information about the network is the Laplacian matrix. This matrix is defined as follows: Let $\mathbf{D}$ be a diagonal matrix where each diagonal entry D_{ii} corresponds to the degree of the vertex i in the graph, i.e. $D_{ii} = \sum_{j=1}^{n} A_{i,j}$, where A is the adjacency matrix of the graph. Then, the Laplacian matrix of the graph is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. It is worth noting that the Laplacian matrix is symmetric and positive semi-definite, and has a zero as eigenvalue, with the associated eigenvector being the vector with all entries equal to one. The main properties the the Laplacian matrix are given in [vL07, section 3]. One of the most important is that the spectrum of $\mathbf{L}$ provides information about the connectivity and the community structure of the graph. The next proposition is a well-known result in graphs literature that can also be found in [vL07].

Proposition 1. *The number of connected components in a graph is equal to the multiplicity of the zero eigenvalue of its Laplacian matrix, and the corresponding eigenvectors are the indicators of these connected components.*

This proposition emphasizes the link between the community structure of a graph and the spectral properties of its adjacency matrix. This will motivate the spectral clustering algorithm presented before in subsection 1.1.5.

1.1.2 Bipartite graphs

In this thesis, our focus will be on bipartite graphs, which are a specific type of graph where the set of vertices, represented by V, is divided into two distinct sets, denoted as V_1 and V_2. More specifically, V can be expressed as the union of V_1 and V_2, and the intersection between V_1 and V_2 is the empty set. The key characteristic of bipartite graphs is that edges in E can only be formed between a vertex in V_1 and a vertex in V_2. We often refer to V_1 as the left hand side of the bipartite graph and V_2 as the right hand side as visualize in figure 1. This structure is commonly used to model interactions between two different types of individuals or entities.

Bipartite networks, as well as graphs, can be fully represented by there adjacency matrix (sometimes called bi-adjacency matrix). Let denote $n = |V_1|$ and $m = |V_2|$ the cardinal of the sets of vertices V_1 and V_2, then the adjacency matrix of the bipartite graph $\mathcal{G} = (V_1 \sqcup V_2, E)$ is the matrix $\mathbf{A}$ of size $n \times m$ such that for every $i \in V_1$ and $j \in V_2$,

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

As before, in more general settings, edges could be labeled by real values, then matrix $\mathbf{A}$ lives in $\mathbb{R}^{n \times m}$, and its coefficients are the edges label (the 0 label means no edge). Notice that $\mathbf{A}$ is no longer automatically symmetric, even if $n = m$, this will be one of the most important issue to tackle along this thesis.

The paper [DG14a] investigates the dynamics between individuals' personality traits and their achievements within the marriage market. The research focuses on an example of a bipartite network, where the two distinct groups of vertices represent men and women. The presence of an edge in this network signifies a marital union between two individuals. By analyzing this bipartite structure, the study aims to uncover potential correlations between individuals' personality traits and their levels of success in finding a spouse. This is the sort of situation where bipartite graphs arise as natural models or objects, that is, when there are really two distinct types of nodes.

The statistical properties of a network dataset can be characterized by the statistical properties of its adjacency matrix $\mathbf{A}$. As such, for our study, both the adjacency matrices and

associated graphs will be considered random. One popular class of random graph models is the stochastic block model.

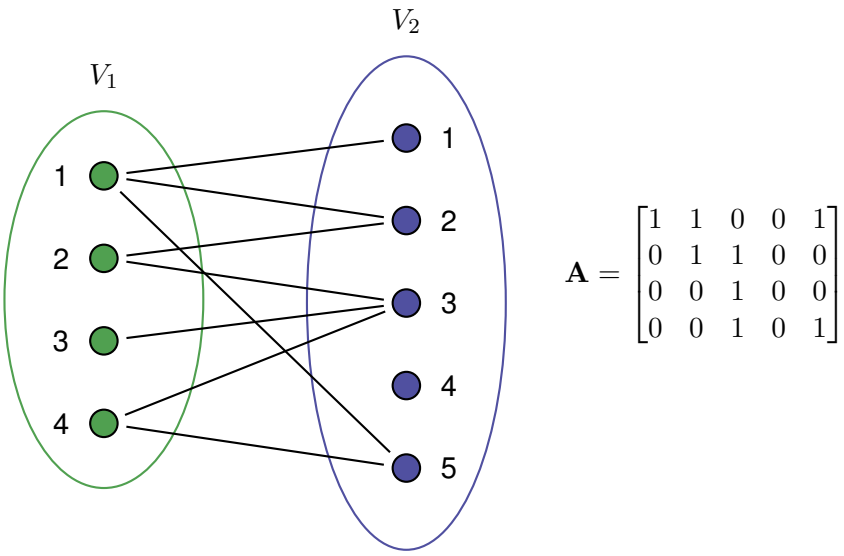


Figure 1: Example of a bipartite graph and its adjacency matrix.

1.1.3 The stochastic block model

Returning to unipartite graphs can help us better understand the stochastic block model (SBM), which is a widely-used generative model for random graphs exhibiting community structure. In this model, subsets of vertices have higher edge densities with each other than with vertices in other subsets. It was first introduced in 1983 by [HLL83] as a mathematical formulation in the field of social networks. The SBM has been an active area of research for at least two decades and has significant applications in statistics, machine learning, and network science. It is commonly used as a benchmark for evaluating community detection algorithms for network data. The stochastic block model takes the following parameters:

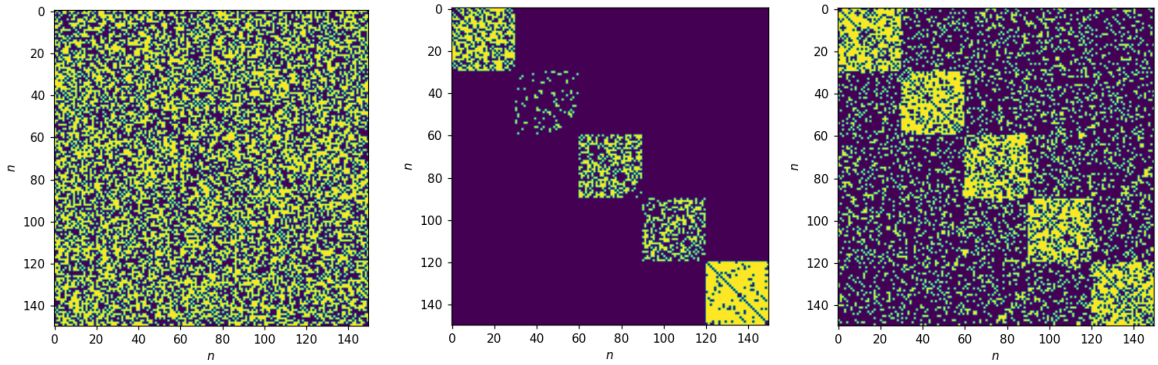
- The number of vertices of the graph, denoted by n.
- The partition of the set of vertices $\{1, \dots, n\}$ into blocks $C_1, \dots, C_K$. The sets C_k of the partition are often called communities or clusters.
- The symmetric matrix $\mathbf{P} \in [0,1]^{K \times K}$ of probabilities of connections between vertices, depending on what clusters the vertices belong to.

To build a graph on a given set of vertices, we need to define the set of edges. For the SBM, the set of edges is independently sampled at random as follow : two vertices $u \in C_k$ and $v \in C_\ell$ are connected with probability $P_{k,\ell}$. In other words, the probability for a vertex u to be connected to an other vertex v depend only on the community to which u and v belong. In term

of adjacency matrix, the SBM corresponds exactly to sample a random symmetric matrix $\mathbf{A}$ with independent entries $A_{i,j}$ drawn from the Bernouilli distribution of parameter $P_{k,\ell}$ if $i \in C_k$ and $j \in C_\ell$. We can give some particular examples of the SBM, that are described bellow and also depicted in Figures 2 and 3.

- When matrix $\mathbf{P}$ has all its entries equal to some $p \in [0, 1]$, the obtained network is an *Erdős–Rényi* graph, where all the edges appears independently and with the same probability p. This case corresponds to $K = 1$, that is there is only one community.
- When the matrix $\mathbf{P}$ is diagonal, it implies that vertices are only connected within their own community, leading to a strong segregation phenomenon. The adjacency matrix in this case would be block diagonal with blocks representing K disconnected communities.
- The specific instance of the SBM, where the diagonal entries of $\mathbf{P}$ are all equal to p, while the off-diagonal entries are equal to q, is called *planted partition model*. In this model, two vertices in the same community are connected with probability p, while two vertices in different communities are connected with probability q. This prototype of the SBM is the most frequently analysed in the literature. It is called an assortative model if $p > q$, and disassortative model, when $p < q$.
- When all the diagonal entries of $\mathbf{P}$ are strictly larger than all off-diagonal entries, the model is called strongly assortative. This means that vertices in the same community are more strongly connected to each other than vertices in different communities.
- When each diagonal entry is strictly larger than the other entries in its own row and column, the model is called weakly assortative. It extends the strongly assortative model. Similar definitions exist for disassortative models, where the inequalities are reversed. These conditions can affect the efficiency of community detection algorithms, with some algorithms performing better on assortative or disassortative block models.

The stochastic block model can be extended to bipartite networks. In contrast to unipartite networks, where a single type of community structure is present, bipartite networks have two distinct types of community structure, one for each side of the bipartite graph. It can model a situation where some workers have to look for some firms to work for. On one side, the workers can be grouped into communities based on factors such as age range or professional skills. On the other side, the firms can be classified into communities based on their line of business. In the statistical applications, the goal is to identify the underlying community structure using the data and not to provide it. The probability matrix associated with a bipartite stochastic block model (BSBM) is not necessarily square, and the notion of assortativity does not apply. A BSBM with two communities per side has been defined and studied in literature, as in [NST22] and [FP16a]. These works also address the community detection problem.



(a) Erdös-Rényi, $p = 0.5$ (b) Diagonal model (c) Planted model $(0.7, 0.2)$

Figure 2: Representation of adjacency matrices for particular examples of the SBM. The central figure represents a diagonal model where the probability matrix $\mathbf{P} = \mathrm{diag}(0.7, 0.1, 0.5, 0.4, 0.9)$ is diagonal.

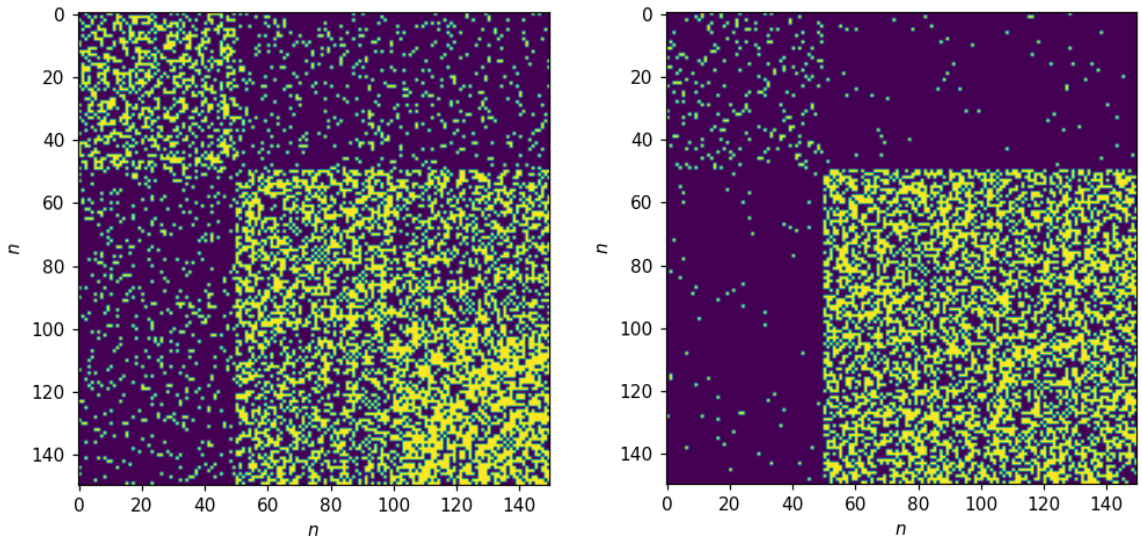
1.1.4 Community detection

Community structures are a frequent occurrence in real-world networks. In social networks, for instance, communities may form around shared interests, locations, occupations, and so on. The presence of communities in a network can also have significant impacts on various processes such as epidemic or rumor spreading. Therefore, detecting and studying communities is crucial for gaining a proper understanding of these processes and how they operate in different settings.

From a statistical standpoint, community detection involves the question of whether we can recover the community structure from an observed network with latent community structure. The observed network could be generated from a fully known, partially known, or unknown SBM. However, to answer this question, we must first define what we mean by "recovery" of the network structure. Typically, there are two types of recovery that researchers consider in the literature.

- *Partial recovery*, which basically means that we are able to asymptotically recover the community structure up to a small fraction of missclustered vertices, that is a large part of the vertices in the network are correctly assigned to their corresponding communities.
- *Exact recovery* is define as a complete asymptotic structure recovery, every vertex in the network is correctly assigned to its corresponding community.

The SBM and specifically the planted partition model discussed in subsection 1.1.3 have been extensively studied in the literature with regards to recovery conditions. [ABH16, page 3] presents a table summarizing some of these conditions in the context of a planted partition model, which depend on the parameters p_n and q_n of the model that vary with the number of vertices n. One notable observation is the following: for certain parameter settings, recovery



(a) Strongly assortative model

(b) Weakly assortative model

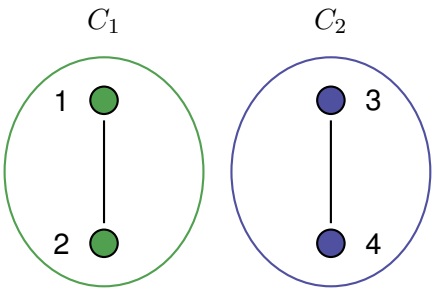
Figure 3: Representation of adjacency matrix for strongly and weakly assortative model. Notice that in figure 3b, there are 3 communities, but two of them are difficult to distinguish. Moreover, we see that the inter-connectivity between communities 2 and 3 located at the bottom right of the figure, is higher than the intra-connectivity within community 1 at the top left. This particular phenomenon cannot occur in a strongly assortative model.

with high probability is achievable, whereas for others, recovery is impossible regardless of the employed algorithm. There are also studies that address the recovery threshold question for bipartite stochastic block models, as seen in [FP16a]. In addition, [NST22, Table 1] presents a summary of recovery conditions and algorithms used for detecting the community structure in bipartite stochastic block models. Numerous algorithms have been developed for both bipartite and unipartite settings, with spectral clustering being one of the most widely used, as discussed in the following subsection.

1.1.5 Spectral clustering

As seen in Proposition 1, the community structure of a unipartite graph is related to the spectrum of the Laplacian matrix. Spectral clustering algorithms are based on this observation to infer the partitioning of the vertices. [DH73] was first suggested using the eigenvectors of adjacency matrices of graphs to find such partitions, and then literature has grown on this topic. A nice overview over the history of spectral clustering can be found in [ST07] or [vL07, section 9].

To give an intuition on spectral clustering, let give a simple example. Consider the Laplacian matrix that corresponds to a graph with two distinct connected components $\{1, 2\}$ and $\{3, 4\}$ and assume that we want to cluster this graph into two communities.



$$\mathbf{L} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

The Laplacian matrix $\mathbf{L}$ has eigenvalues 0 and 2 with both multiplicity 2, and normalized eigenvectors are respectively given by

$$\boldsymbol{u}_1 = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{u}_3 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{u}_4 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}.$$

Let consider the matrix $\mathbf{U} \in \mathbb{R}^{4\times 2}$ formed by the two first eigenvectors. Note that the first two rows are identical, as are the last two rows. Then we naturally group those rows indices together, which give the desired clustering. If we add the third eigenvector in matrix $\mathbf{U}$, we see that row 1 is closer to row 3 than to row 2 (in the sens of the euclidean distance), which leads to miss clustering. This example may be too simple to justify it will works with larger networks. Consider another example with the graph representation given in figure 4. The first

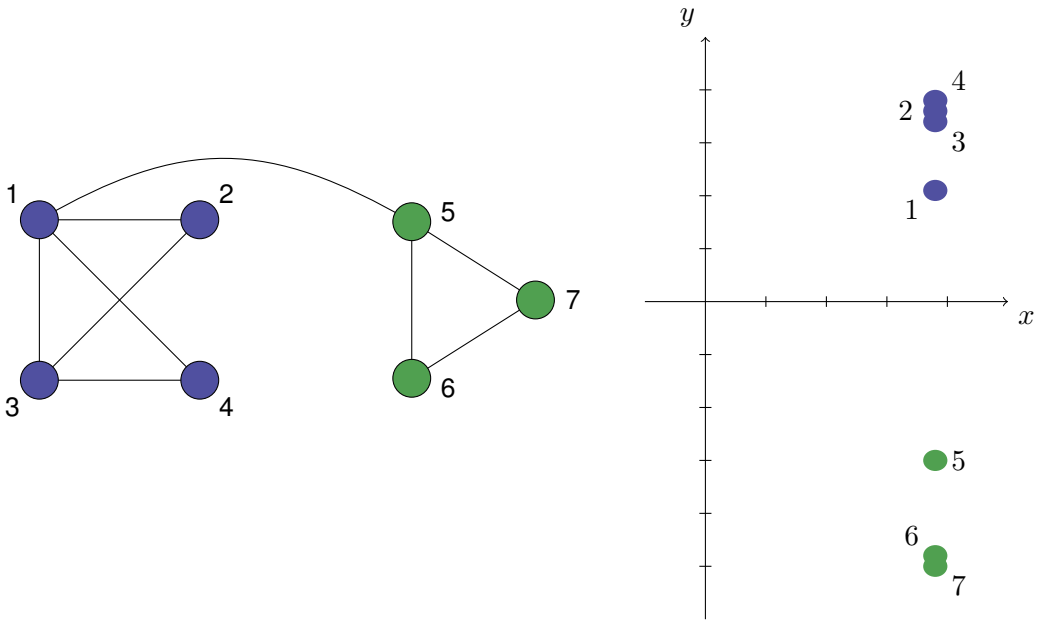


Figure 4: The left figure represents the graph considered, divided into two communities. The right figure is a visualization of points in $\mathbb{R}^2$ represented by the columns of matrix $\mathbf{U}$. To make points 2, 3, 4, as well as 6, and 7 distinguishable, we intentionally perturbed their coordinates. The clusters clearly appear on this figure.

two eigenvectors are given by the matrix rows and plotted in figure 4

$$\mathbf{U} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} \approx \begin{bmatrix} 0.38 & 0.38 & 0.38 & 0.38 & 0.38 & 0.38 & 0.38 \\ 0.21 & 0.36 & 0.36 & 0.36 & -0.30 & -0.49 & -0.49 \end{bmatrix}.$$

We observe that columns 2, 3, and 4 are identical, as well as columns 6 and 7. Moreover, column 1 is closer to column 2, 3, and 4 than to column 6 and 7 and so on. Based on this observation, if we want to cluster the graph into two communities, we will naturally group vertices 1, 2, 3, and 4 together, and vertices 5, 6, and 7 together. This procedure consisting in grouping the closest rows (or collumns) together is called k-means clustering. On this example, adding the third eigenvector in matrix $\mathbf{U}$ does not affect the clustering, but when we consider the five first eigenvectors, the clusters are not recovered. Obviously, if we consider only the first eigenvector, it gives no information about the community structure of the graph. This make us understand that the good choice of the number of eigenvectors to consider is exactly the number of clusters we want in our graph.

So this approach generalizes to the case of larger networks: we compute the first K eigenvectors of the Laplacian, form the matrix $\mathbf{U}$, and apply the k-means algorithm to group together rows of $\mathbf{U}$ that are the "closest". This procedure is described in algorithm 1.

Algorithm 1 Clustering algorithm

Require: $\mathbf{A} \in \mathbb{R}^{n \times n}$ the adjacency matrix of a graph, $K \in \mathbb{N}$ the number of clusters.
Ensure: $C_1, \dots, C_K$ the partition of $[n]$ into K-clusters.
1: Compute $\mathbf{L}$ the Laplacian matrix associated to $\mathbf{A}$.
2: Compute the K-first eigenvectors $\boldsymbol{u}_1, \dots \boldsymbol{u}_K$ of $\mathbf{L}$ where $0 = \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_K$ are the associated eigenvalues ordered in the increasing order.
3: Let $\mathbf{U} \in \mathbb{R}^{n \times K}$ be the matrix whose columns are the $\boldsymbol{u}_i$'s.
4: Cluster the rows of $\mathbf{U}$ with the k-means algorithm into K-subsets $C_1, \dots, C_K$ of $[n]$.

There are also clustering algorithms for bipartite networks, which operate similarly to those for networks. These algorithms work directly on the adjacency matrix, and distinguish between the left and right eigenvectors to recover the clusters of the left and right hand sides respectively (see algorithm 2 for an example of clustering algorithm to recover the left clusters). Interested readers are referred to [ZA19a], where spectral clustering algorithms for community detection under a general bipartite stochastic block model are considered.

Algorithm 2 Clustering algorithm for bipartite networks (left-hand side)

Require: $\mathbf{A} \in \mathbb{R}^{n \times m}$ the adjacency matrix of a bipartite graph, $K \in \mathbb{N}$ the number of clusters for the left-hand side.
Ensure: $C_1, \dots, C_K$ the partition of $[n]$ into K-clusters.
1: Compute the K-first left-singular vectors $\boldsymbol{u}_1, \dots \boldsymbol{u}_K$ of $\mathbf{A}$ where $\sigma_1 \leqslant \sigma_2 \leqslant \cdots \leqslant \sigma_K$ are the associated singular values ordered in the increasing order.
2: Let $\mathbf{U} \in \mathbb{R}^{n \times K}$ be the matrix whose columns are the $\boldsymbol{u}_i$'s.
3: Cluster the rows of $\mathbf{U}$ with the k-means algorithm into K-subsets $C_1, \dots, C_K$ of $[n]$.

1.2 Definition of the problem

Preliminary considerations In the field of economics, network datasets are frequently used to model markets or interactions between different entities or individuals. As a result, there is a vast literature on the topic of economic networks and link formation models, as evidenced by the works of [Gra17, Gra20, DG14b, JW96, Dze19]. Both unipartite and bipartite networks can be relevant for economic modeling. For example, international trade exchange can be modeled using a unipartite graph where links represent the existence of trade exchange. On the other hand, bipartite graphs are more appropriate for modeling consumer purchases of products or workers hiring in a firm.

When modeling network formation, there are typically two types of variables to consider: observable and unobservable.

- Observable variables are characteristics that we can access or compute from the data we observe. For instance, in the context of international trade, we may have information about the size, GDP or the geographic location of each country.
- On the other hand, unobservable variables are latent variables that we cannot observe or calculate directly. In the case of the network of workers and firms, examples of unobservable variables include for instance the sympathy of a worker, which may influence the hiring process, or the attractivity (the good working atmosphere) of a firm, which may influence the workers' choices.

[Gra17] propose an econometric model of network formation for unipartite graph with both observed and unobserved variables. In contrast [Gra20] presents a logistic model for bipartite network with only observed variables.

Now comes the question of assumptions about the formation of links. The most common assumption is that all links are formed independently, which is more or less realistic or sometimes even irrelevant depending on the modeling situation, but it makes the mathematical problems easier to deal with. A more realistic and weaker assumption is described in [DDG21a], that talk about exchangeability. Roughly speaking, it means that we can permute the label of the vertices without changing the distribution of the edges, or equivalently, we can apply any permutation of rows and columns indices to the adjacency matrix without changing its distribution.

According to [Ald81, Theorem 1.4], a bipartite graph with row and column exchangeable adjacency matrix $\mathbf{A}$ can be represented by a function $g^* : [0,1]^4 \to \mathbb{R}$ and independent uniformly distributed random variables α, U_i, V_j, ξ_{ij} in $[0,1]$ in the sens that

$$\mathbf{A} \overset{(d)}{=} \left(g^*(\alpha, U_i, V_j, \xi_{ij}); i \in [n], j \in [m]\right)$$

Here, the random variables U_i, V_j and $\xi_{i,j}$ correspond to the aforementioned unobserved variables. If additional observable feature vectors $\boldsymbol{X}_{i,j}$ are available for each node pair (i,j),

then an extended model can be considered, which is defined by $g^*(\boldsymbol{X}_{i,j}, \alpha, U_i, V_j, \xi_{i,j})$. We aim to estimate the function g^* using the graph it generates, but this task can be challenging. To simplify the problem, we propose to consider only the unobserved variables by removing all observable variables. In a first step, we assume that the edges are formed independently conditionally to the unobserved variables. We formally define the problem here after. In a second step, we want to relax the independence assumption. Indeed, if we consider the worker-firm network problem, a worker must choose one and only one firm to work for, thus, links are not independent anymore.

Mathematical definition of the main problem Let n and m be two positive integers assumed to be large, and $\mathbf{H}$ be an $n \times m$ random matrix with real entries $H_{i,j}$. The entries of the matrix $\mathbf{H}$ can be seen as the edge labels of a bipartite graph. We assume that the distribution of this matrix $\mathbf{H}$ satisfies the following condition.

Assumption 1 (Full independence)**.** There is a function $W^* : [0,1]^2 \to \mathbb{R}$, called the graphon, and two random vectors $\boldsymbol{U} = (U_1, \dots, U_n)$ and $\boldsymbol{V} = (V_1, \dots, V_m)$ such that

A 1.1 $U_1, \dots, U_n, V_1, \dots, V_m$ are independent and drawn from the uniform distribution $\mathcal{U}([0,1])$.

A 1.2 conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the entries $H_{i,j}$ are independent and $\mathbb{E}[H_{i,j}|\boldsymbol{U}, \boldsymbol{V}] = W^*(U_i, V_j)$.

Assumption 1 should be understood as follows. Each vertex on the left-hand side of the bipartite graph is assigned an unobserved variable U_i, and similarly for the right-hand side with the variables V_j. Furthermore, assuming we know the variables U_i and V_j, we posit that the entries of the adjacency matrix $\mathbf{H}$ are independent. To illustrate, let's consider an example where they are drawn from a Bernoulli distribution with parameter $W^(U_i, V_j)$—meaning an edge between i and j is present with a probability of $W^(U_i, V_j)$. More general distributions will be considered later, allowing for edge labels.

We aim to investigate the minimax risk of estimating the graphon W^* from the observation of $\mathbf{H}$, and demonstrate how it relies on crucial parameters of the problem. While the dimensions of the matrix n and m are among these parameters, we also explore the impact of the smoothness of W^*, the degree of "sparsity" in interactions (represented by ρ), and the noise level (represented by σ). To be more accurate, σ and ρ are positive real numbers such that

$$\|W^*\|_\infty = \sup_{u,v \in [0,1]} |W^*(u,v)| \leqslant \rho$$
$$\text{and} \qquad \mathbf{Var}[H_{i,j}|U_i, V_j] \leqslant \sigma^2 \text{ a.s.}, \quad \forall i \in [n], \forall j \in [m].$$

To fix the idea, let assume that W^* is piecewise constant with respect to a partition of the unit square in axis-aligned rectangles, and denote $\mathbf{\Theta}^* \in [0,1]^{n \times m}$ the random matrix given by

$$\Theta^*_{i,j} = W^*(U_i, V_j).$$

Notice that $\mathbf{\Theta}^*$ is constant by block up to some permutations of rows and columns. Suppose moreover that conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, entries $H_{i,j}$ are independently drawn from the Bernoulli distribution with parameter $\Theta^*_{i,j}$. This scenario corresponds to the bipartite stochastic block model discussed in the previous subsection 1.1.3, where the associated probability matrix $\mathbf{\Theta}^*$ is unknown. Thus, the problem of estimating W^* is equivalent to the community detection problem for bipartite networks, as discussed in the previous subsections, where we seek to estimate the clusters and the probability matrix. In addition to the Bernoulli distribution, we aim to provide a risk upper bound for our estimation method for more general distributions. In this thesis, we will also consider non parametric framework: the class of α-Hölder regular graphons as another form of graphon regularity. To do so, we aim to approximate Hölder graphons by piecewise constant graphons (see figure 5).

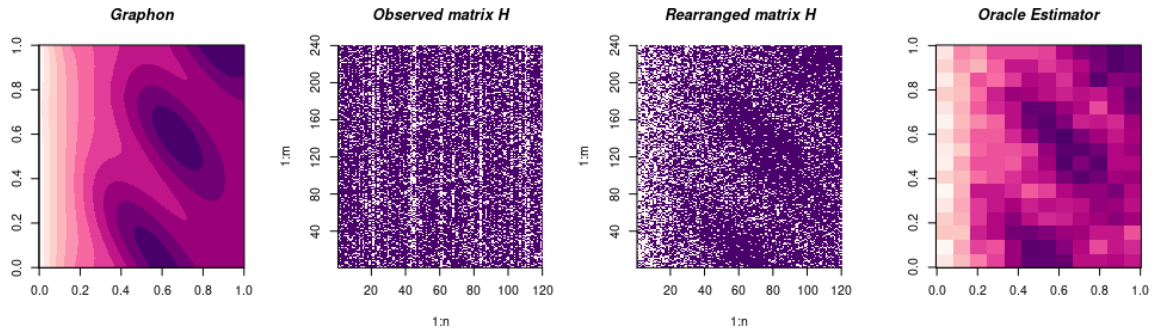


Figure 5: An illustration of the graphon problem. The leftmost graph represents the unknown graphon W^*. The second leftmost graph is the adjacency matrix observed in the graph where the links are made according to the Bernoulli model. The third graph is the adjacency matrix that would be obtained after a rearrangement of the rows and columns if we had access to the latent variables. The rightmost graph represents the histogram estimator obtained from the rearranged adjacency matrix. Our goal is to design an estimator which is nearly as good as the oracle, without having access to the latent variables.

Relaxation of the independence assumption The previous setting, which assumes independence between edges, may not be suitable for modeling certain common situations encountered in practice. One such example is the worker-firm network, where the first set of vertices represents workers and the second set represents firms. A worker is connected to a firm if he is hired by that firm. In this scenario, it is reasonable to assume that each worker is hired by at most one firm, resulting in a maximum degree of 1 for each vertex in the first set. The idea, then, is to relax the independence assumption regarding link formation. We also consider only unobserved latent variables in this model, and assume that the adjacency matrix $\mathbf{H}$ now lives in $[0,1]^{n\times m}$ and satisfy the following statement.

Assumption 2 (Relaxed independence)**.** We consider a function $W^* : [0,1]^2 \to [0, +\infty[$ called the graphon and two random vectors $\boldsymbol{U} = (U_1, \dots, U_n)$ and $\boldsymbol{V} = (V_1, \dots, V_m)$ that satisfy

A 2.1 $U_1, \dots, U_n, V_1, \dots, V_m$ are independent and drawn from the uniform distribution on $[0,1]$.

A 2.2 Conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the rows of the matrix $\mathbf{H}$ are independent.

A 2.3 Each row of $\mathbf{H}$ sum to one and

$$\mathbb{E}[H_{i,j}|\boldsymbol{U}, \boldsymbol{V}] = \frac{W^*(U_i, V_j)}{\sum_{\ell=1}^{m} W^*(U_i, V_\ell)}.$$

Assumption **A 2.1** is the same as **A 1.1**, we consider that the unobserved variables assigned for each vertex are independent. However, assumption **A 2.2** relaxes the independence assumption on the links formed by individuals on the right-hand side. Instead, it only requires that the rows of $\mathbf{H}$ are conditionally independent given $(\boldsymbol{U}, \boldsymbol{V})$. In other words, while the edges formed by distinct individuals on the right-hand side are independent, the links formed by a single individual are not necessarily independent.

Our goal is twofold. First, we aim to estimate the mean matrix of $\mathbf{H}$ and provide a risk bound for our estimation method. In this part, we do not consider part **A 2.3** of assumption 2 that could be restrictive, but instead replace it with the following assumption:

$$\sum_{j=1}^{m} H_{i,j} \leqslant \rho_{\mathbf{H}} \quad \forall i = 1, \dots, m \qquad \textbf{(A 2.3} \text{ (bis))}$$

that is, the sum of each row of $\mathbf{H}$ is smaller than a positive parameter $\rho_{\mathbf{H}}$. We will often consider $\rho_{\mathbf{H}} = 1$. Our goal is to analyze how the risk of our estimator behaves in relation to the model parameters, specifically n and m, the size of our dataset, as well as the parameters $\rho_{\mathbf{H}}$, representing the row sum constraint, and $\rho_{\mathbf{\Sigma}}$ the noise level, which now satisfies [1]

$$\|\mathbf{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\mathbf{\Sigma}} \quad \forall i = 1, \dots, n \tag{1.1}$$

where $\mathbf{\Sigma}_i = \mathbb{E}[\boldsymbol{H}_i \boldsymbol{H}_i^\top] - \mathbb{E}[\boldsymbol{H}_i]\mathbb{E}[\boldsymbol{H}_i]^\top$ is the covariance matrix of the ith row of $\mathbf{H}$. It will be common to assume that the matrix $\mathbf{\Theta}^* = \mathbb{E}[\mathbf{H}]$ has rows whose sum is bounded by $\rho_{\mathbf{\Theta}} > 0$. This parameter may also appear in the upper bounds of the risk. Notice that $\rho_{\mathbf{\Theta}}$ could be much smaller than $\rho_{\mathbf{H}}$. For example, if $\boldsymbol{H}_i$ is multinomial with parameters $(1, m, (1/m, \dots, 1/m))$, then $\rho_{\mathbf{H}} = 1$ and $\rho_{\mathbf{\Theta}} = 1/m$. The second part of our work revolves around estimating the graphon W^* under the assumption 2 and condition (1.1). Furthermore, we focus exclusively on the class of piecewise constant graphons for this estimation task.

1.3 Prior work

The field of economics extensively utilizes network datasets to model markets and interactions among various entities and individuals. Consequently, there is a wide range of literature focusing on economic networks and models for link formation. This is evident in the works of

[1] $\|\mathbf{\Sigma}\|_{\mathrm{op}}$ refers to the operator norm of a square matrix.

[Gra17, Gra20, DG14b, JW96, Dze19, dPRST18]. Economic modeling can encompass both unipartite networks as evidenced by [OW14b, Section 4], and bipartite networks as in [DG14b]. Both types of graph are relevant in this context.

For the past two decades, there has been significant research activity in the statistical analysis of matrix and network data using block models similar to those considered in this work. One prominent example is the stochastic block model, which was originally introduced by [HLL83] and has since become one of the most extensively studied latent structures for network data. Early references include [NS01] and [GN03]. A key focus of research in this field has been community detection, which aims to uncover the underlying block structure within networks. Numerous studies, such as [ZLZ12, CRV15, LR15, Lei16, ZZ16, WB17, CLX18, XJL20], have delved into this problem. Several studies, such as the one conducted by [CDP12], have dealt with the estimation of parameters in a stochastic block model using accessible techniques like variational inference. Notably, [GK21] presents a groundbreaking contribution by introducing the first minimax-optimal and tractable estimator for parameter estimation in the stochastic block model, especially when dealing with missing links. Additionally, [LM19] underscores the efficacy of variational methods in approximating the maximum likelihood estimator for dynamic stochastic block models.

Additionally, literature reviews by [GMZZ17, Abb18] provide comprehensive overviews of the subject. It is important to note that the majority of these studies have primarily focused on unipartite graphs with binary or discrete edge labels. The optimality of their approaches has largely revolved around determining the minimum separation rate between community parameters that allows for consistent recovery. While much attention has been given to unipartite block models, similar problems for bipartite block models have also been investigated. Studies by [FPV15, FP16b, Neu18, ZA19b, ZA20, CLC$^+$21, NST22] have explored these issues, shedding light on the unique characteristics of bipartite networks.

In contrast to the papers mentioned earlier, our focus is on the optimality of estimation error in a model that encompasses bipartite graphs with real-valued edge labels. Consistency of graphon estimators has been extensively studied in [ACC13, WO13, OW14a]. Moreover, minimax-rate optimality of the least squares estimator has been established in [GLZ15, GLMZ16, KTV17, KV19], with additional insights provided in the survey article [GM21]. To better position our contributions within the existing state-of-the-art, it is beneficial to provide a brief overview of the content covered in these papers.

In [GLZ15], the authors focused on binary observations $H_{i,j}$ and considered the dense case $\rho \asymp 1$. They derived minimax rates of estimation for both piecewise constant and Hölder continuous graphons. Building upon these results, [GLMZ16] extended the analysis to matrices $\mathbf{H}$ with sub-Gaussian entries, some of which may be missing completely at random. However, as discussed in subsection 2.3.1, their results are sub-optimal in certain cases concerning the noise variance, such as when edge labels are drawn from the binomial distribution. For unipartite graphs with binary edge labels, [KTV17] established the minimax-optimal rates of

estimation for sparsely connected graphs where $\rho \ll 1$. While [GLZ15, GLMZ16] measured the estimation error using the normalized Frobenius norm of the difference between the estimated matrix and the true matrix, [KTV17] also considered the $\mathbb{L}_2$-distance between the equivalence classes of graphons. In [KV19], minimax optimal rates of graphon estimation in the cut distance were established for unipartite graphs with binary observations. These findings contribute to the current understanding of optimality in various estimation scenarios.

The problems investigated in the initial part of our work exhibit connections with recent developments in econometrics. Specifically, as previously mentionned, according to [Ald81, Theorem 1.4], if the matrix $\mathbf{H}$ satisfies row and column exchangeability, then there exists a function $g^* : [0,1]^4 \to \mathbb{R}$ and independent uniformly distributed random variables α, U_i, V_j, ξ_{ij} in $[0,1]$ such that the random matrices $\mathbf{H}$ and $\big(g^*(\alpha, U_i, V_j, \xi_{ij}); i \in [n], j \in [m]\big)$ follow the same distribution. The problem addressed in chapter 2 is equivalent to estimating the random function $W^(u,v) = \int_0^1 g^*(\alpha, u, v, z), dz$ based on observations $g^*(\alpha, U_i, V_j, \xi_{ij}); i \in [n], j \in [m]$, without assuming a specific parametric form for g^*. Furthermore, if we have additional feature vectors $\boldsymbol{X}_{i,j}$ associated with each node pair (i,j), we can consider an extended model defined by $g^*(\boldsymbol{X}_{i,j}, \alpha, U_i, V_j, \xi_{i,j})$. This approach is employed in [Gra17], where a particular parametric form $g^*(\boldsymbol{x}, \alpha, u, v, z) = \mathbb{1}\big(\boldsymbol{x}^\top \boldsymbol{\beta} + u + v + \log(z/(1-z))\big)$ is considered. In such a parametric setting, the parameter of interest is the vector $\boldsymbol{\beta}$, measuring the homophily of the graph (the tendency of individuals to form connections with those like them-selves). Additionally, in [Gra20], the assumption is made that the regression function $\int_{[0,1]^3} g^*(\boldsymbol{x}, \alpha, u, v, z), du, dv, dz$ follows a parametric form $\exp(a + \boldsymbol{x}^\top \boldsymbol{\beta})/(1 + \exp(a + \boldsymbol{x}^\top \boldsymbol{\beta}))$, and the estimation problem for the vector $(a, \boldsymbol{\beta})$ is studied. Asymptotic results, such as the law of large numbers and central limit theorem, have been established for exchangeable arrays in [DDG21b].

1.4 Contributions

We provide a summary of the main contributions of this thesis in four subsections.

- The first subsection focuses on the mean estimation problem, which is crucial in both independence contexts described in assumptions 1 and 2. This step is significant and interesting on its own.
- The second subsection addresses the graphon estimation problem, building upon the estimation procedure derived from the first step.
- In the third subsection, we establish lower bounds on the worst-case risk for any graphon estimator over the set of piecewise constant graphons, under assumption of full independence links. These lower bounds are applicable when the matrices have entries with a Binomial conditional distribution given latent variables. Remarkably, in most cases, these lower bounds are of the same order as the upper bounds obtained for the least-squares

estimator.

- Finally, in the fourth subsection, we present an adaptation of Lloyd's algorithm of alternating minimization, incorporating a convex relaxation step, to our specific setting. This adaptation allows us to obtain a computationally tractable approximation of the least squares estimator, and some simulations base on synthetic data, only for the full independence setting.

For the convenience of theorem statements, we will often adopt the symmetric framework, where both sides of the bipartite graphs have an equal number of vertices, communities, etc.

1.4.1 Estimation of the mean matrix

Full independence assumption On the way to estimating the graphon W^*, an important intermediate step will consist in estimating the matrix $\mathbf{\Theta}^* = W^*(U_i, V_j)$. The estimation of this matrix is of interest on its own. We perform this task by solving the least squares problem over the set of constant-by-block matrices, with blocks generated by partitions of the sets of rows and the columns of the matrix $\mathbf{H}$. It will be further shown that the method of aggregation by exponential weights can be used to ensure adaptivity to the number of blocks. Under the condition that the graphon is piecewise constant or α-smooth in the sense of Hölder smoothness, we establish risk bounds for the graphon estimator derived from the estimator of $\mathbf{\Theta}^*$. These risk bounds are nonasymptotic, and shown to be rate optimal in the minimax sense for a broad range of regimes.

The least square estimators of $\mathbf{\Theta}^*$ are define as the best approximation of $\mathbf{H}$ by a constant by block matrix. To be more accurate

$$\widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K,L] \in \arg\min_{\mathbf{\Theta}\in\mathcal{T}^{K,L}_{n_0,m_0}} \|\mathbf{H}-\mathbf{\Theta}\|_{\mathsf{F}}^2. \tag{1.2}$$

Here, $\mathcal{T}^{K,L}_{n_0,m_0}$ represents a set of constant by block matrices (up to some permutations of rows and columns) with $K \times L$ blocks. Parameters $n_0 \geqslant 1$, and $m_0 \geqslant 1$ refer to the minimal number of entries in each block. We derive risk bounds for estimators of $\mathbf{\Theta}^*$. For the purpose of our analysis, we consider $\mathbf{\Theta}^*$ as a deterministic matrix, allowing us to assume independence of $H_{i,j}$ rather than conditional independence given $\boldsymbol{U}$ and $\boldsymbol{V}$. Furthermore, for the sake of simplicity, in the statement of the next theorem, we make the assumption of symmetry, where $n = m$, $K = L$, and $n_0 = m_0$.

Theorem 1. *Let n, n_0, K be positive integers such that $K \geqslant 2$, $3 \leqslant n_0 \leqslant n$. Let $\mathbf{H}$ be an $n \times n$ random matrix with independent entries satisfying $\mathbb{E}[H_{ij}] \in [0, \rho]$ for every $i, j \in [n]$ and some $\rho > 0$. In addition, assume that the random variables $(H_{ij} - \mathbb{E}[H_{ij}])$ satisfy the (σ^2, b)-Bernstein condition [2]. Then, the least squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ of the mean matrix $\mathbf{\Theta}^* = \mathbb{E}[\mathbf{H}]$, defined by*

[2]We say that a zero-mean random variable ζ satisfies the (a, b)-Bernstein condition, if we have $\mathbb{E}[e^{\lambda\zeta}] \leqslant$

(1.2), *satisfies the exact oracle inequality*

$$\frac{1}{n}\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2} \leqslant \inf_{\mathbf{\Theta}\in\mathcal{T}_{n_0}^K} \frac{1}{n}\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}} + \big(25\sigma^2 + 4b\rho\big)^{1/2}\Big(\frac{3K^2}{n^2} + \frac{2\log K}{n}\Big)^{1/2},$$

provided that $\psi_n(n_0) := \frac{6}{n_0}\log(en/n_0) \leqslant (\sigma/b)^2$.

In table 1.1, we provide four main examples that illustrate the consequences of Theorem 1 in the non symmetric case for common distributions of $H_{i,j}$. It is worth noting that in the

Model	Definition	(σ^2, b)	Upper-bound
Bernoulli	$H_{i,j} \sim \mathcal{B}(\Theta^*_{i,j})$	$(\rho, 1/3)$	$9\sqrt{\rho}\Big(\frac{KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2}$
Binomial	$NH_{i,j} \sim \mathcal{B}(N, \Theta^*_{i,j})$	$(\rho/N, 1/3N)$	$9\sqrt{\rho}\Big(\frac{KL}{Nnm} + \frac{\log K}{Nm} + \frac{\log L}{Nn}\Big)^{1/2}$
Poisson	$TH_{i,j} \sim \mathcal{P}(T\Theta^*_{i,j})$	$(\rho/T, 1/3T)$	$9\sqrt{\rho}\Big(\frac{KL}{Tnm} + \frac{\log K}{Tm} + \frac{\log L}{Tn}\Big)^{1/2}$
Sub-Gaussian	$\mathbb{E}[e^{\lambda H_{i,j}}] \leqslant e^{\sigma^2\lambda^2}$	$(\sigma^2, 0)$	$5\sigma\Big(\frac{KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2}$

Table 1.1: Here is a summary of the second term obtained for upper bounds in the non symmetric version of Theorem 1 for specific examples of (σ^2, b)-Bernstein distributions. In all cases, we make the assumption that $\Theta^*_{i,j} \leqslant \rho$, except for the sub-gaussian model.

symmetric case, we retrieve the upper-bound derived for unipartite graphs as described in [KTV17]. This indicates that the results obtained in the current context extend and align with the findings in the unipartite graph setting.

The least squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ exhibits a blockwise constant pattern with KL blocks. The choice of KL as the number of blocks is a hyperparameter of the method. However, if the true matrix $\mathbf{\Theta}^*$ significantly deviates from being blockwise constant on KL blocks, the estimation quality of $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K, L]$ may deteriorate due to a substantial bias. To mitigate this bias, one approach is to compute the least squares estimator for multiple values of K, L, n_0, and m_0, and then aggregate these estimators. By doing so, the bias can be reduced and the overall estimation performance can be improved. We also provide finite sample risk bound for this type of aggregate estimator. Finally,the mathematical results can be readily adapted to the case of missing observations, where certain values of the matrix $H_{i,j}$ are not observed.

Relaxed independence assumption In the case of relaxed independence assumption, we are able to derive a comparable upper-bound. However, it is important to note that the distributions applying in Theorem 2 are not as general as those assumed for full independence in the previous framework. Although the scope may be narrower, these distributions still provide

$\exp\big\{\frac{\lambda^2 a}{2(1-b|\lambda|)}\big\}$ provided that $|\lambda| \leqslant 1/b$.

valuable insights and results for our analysis. Once again, we state the next theorem only in the symmetric framework.

Theorem 2. *Let n, n_0, K be positive integers such that $K \geqslant 2$, $1 \leqslant n_0 \leqslant n$. Let $\mathbf{H} \in [0,1]^{n \times n}$ be an $n \times n$ random matrix with independent rows such that each row sum to one and has a covariance matrix $\mathbf{\Sigma}_i$ satisfying $\|\mathbf{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\mathbf{\Sigma}}$. We also assume that $\|\mathbf{\Theta}^*\|_\infty \leqslant \rho_\infty$ [3]. The least-squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ defined by (1.2) satisfies the exact oracle inequality*

$$\frac{1}{n}\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2} \leqslant \inf_{\mathbf{\Theta} \in \mathcal{T}} \frac{1}{n}\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}} + \big(48\rho_{\mathbf{\Sigma}} + 6\rho_\infty\big)^{1/2}\Big(\frac{3K^2}{n^2} + \frac{2\log K}{n}\Big)^{1/2}$$

provided that $\psi_n(n_0) = \frac{2\log(ne/n_0)}{n_0} \leqslant \rho_{\mathbf{\Sigma}}$.

This result contains the case described previously where the vector $\boldsymbol{H}_i$ has only one entry equals to 1, and the others are null, which models a matching where individuals from the left hand size have to choose one and only one item from the right hand size, as in the worker-firm network.

1.4.2 Estimation of the graphon

Full independence assumption The illustration in figure 5 highlights the impact of missing knowledge of latent variables on the graphon estimation problem. It demonstrates that the rearranged adjacency matrix, obtained if the latent variables were known, provides significantly more information about the true graphon W^* compared to the original adjacency matrix. When the latent variables $\boldsymbol{U}$ and $\boldsymbol{V}$ are unknown, the graphon W^* becomes unidentifiable. We define equivalence between two graphons W and W' if there exist two bijections $\tau_1 : [0,1] \to [0,1]$ and $\tau_2 : [0,1] \to [0,1]$ that preserve the Lebesgue measure, such that $W = W' \circ (\tau_1 \otimes \tau_2)$ [4]. It can be observed that two matrices $\mathbf{H}$ generated by equivalent graphons W^* and $\bar{W}^*$ have the same distribution. Therefore, the best we can do is to estimate the equivalence class containing W^*. This motivates the use of the (pseudo)-distance employed in this work to evaluate the quality of an estimator $\widehat{W}$ of W^*, as follows:

$$\begin{aligned}\delta(\widehat{W}, W^*) &= \inf_{\tau_1,\tau_2 \in \mathcal{M}} \Big(\iint_{[0,1]^2} \big|\widehat{W}(\tau_1(u), \tau_2(v)) - W^*(u,v)\big|^2 du\, dv \Big)^{1/2} \\ &= \inf_{\tau_1,\tau_2 \in \mathcal{M}} \|\widehat{W} \circ (\tau_1 \otimes \tau_2) - W^*\|_{\mathbb{L}^2}\end{aligned}$$

where $\mathcal{M}$ is the set of all automorphisms $\tau : [0,1] \to [0,1]$ such that τ and τ^{-1} are measurable, and τ preserves the Lebesgue measure in the sense that $\lambda(\tau^{-1}(B)) = \lambda(B)$ for every Borel-set $B \subset [0,1]$.

[3] $\|\mathbf{A}\|_\infty = \max_{i,j} A_{i,j}$ denotes the infinite norm of a matrix.

[4] We use notation $\tau_1 \otimes \tau_2$ for the function from $[0,1]^2$ to $[0,1]^2$ defined by $(\tau_1 \otimes \tau_2)(u,v) = (\tau_1(u), \tau_2(v))$.

After estimating the matrix $\mathbf{\Theta}^*$ and selecting a distance measure for graphon quality assessment, the next step involves designing an estimator for the graphon W^*. To achieve this, we associate a graphon $W_{\mathbf{\Theta}}$ with any $n \times m$ matrix $\mathbf{\Theta}$, where $W_{\mathbf{\Theta}} : [0,1]^2 \to [0,1]$ is defined as a constant function on each rectangle $I_i \times J_j = \left[\frac{i-1}{n}, \frac{i}{n}\right) \times \left[\frac{j-1}{m}, \frac{j}{m}\right)$ for $(i,j) \in [n] \times [m]$:

$$W_{\mathbf{\Theta}}(u,v) = \Theta_{i,j}, \quad \text{for all} \quad (u,v) \in I_i \times J_j.$$

In the forthcoming theorem, we will analyze the estimator $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}$. As mentioned earlier, we will classify W^* into two categories based on its regularity: the class of piecewise constant graphons and the class $\mathbb{H}_{\alpha,\mathcal{L}}$ [5] of α-Hölder graphons. The statement presented here specifically addresses the simplified symmetric case. However, it is worth noting that we have also obtained results for the asymmetric case, which are discussed in detail in Chapter 2. A summary of these results can be found in table 1.2.

Theorem 3. *Let $\mathbf{H}$ be a $n \times n$ random matrix satisfying Assumption 1 with some graphon $W^* : [0,1]^2 \to [0,\rho]$. Assume that for some constant $\sigma > 0$, conditionally to $\boldsymbol{U}, \boldsymbol{V}$, the random variables $(H_{ij} - \mathbb{E}[H_{ij}|\boldsymbol{U}, \boldsymbol{V}])$ satisfy the (σ^2, b)-Bernstein condition.*

1. *Assume that the graphon W^* is K-piecewise constant, meaning that for some integer $K \geqslant 2$ and for $0 = a_0 < \ldots < a_K = 1$ such that*

$$\Delta^{(K)} := \min_{k \in [K]} |a_k - a_{k-1}| \geqslant \frac{8 \log(nK)}{n}$$

the function W^* is constant on each rectangle $[a_{k-1}, a_k[^2$. Then, the estimator $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}$ with $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0}[K]$ defined by (1.2) satisfies

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant \left(27\sigma^2 + 4b\rho\right)^{1/2} \left(\frac{3K^2}{n^2} + \frac{2 \log K}{n}\right)^{1/2} + \rho \left(\frac{2K}{n}\right)^{1/4}, \tag{1.3}$$

provided that $\psi_n(\Delta^{(K)}) = \frac{12 \log(2e/\Delta^{(K)})}{n\Delta^{(K)}} \leqslant (\sigma/b)^2$.

2. *Assume that the graphon W^* is α-Hölder continuous, meaning that $W^* \in \mathbb{H}_{\alpha,\mathcal{L}}$ for some $\alpha \in (0,1]$ and $\mathcal{L} > 0$. Assume that*

$$\frac{n^{2\alpha}}{\log^4(2n)} \geqslant \mathcal{L}^2 \frac{(4b/\sigma)^{4(\alpha+1)} \vee 3}{(25\sigma^2 + 4b\rho)}. \tag{1.4}$$

Let $\beta = \alpha/(2\alpha+2)$. Then, there is a choice of K, n_0 such that the least squares estimator $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}$ with $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0}[K]$ satisfies

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant 6\mathcal{L}^{1-2\beta} \left(\frac{25\sigma^2 + 4b\rho}{3n^2}\right)^{\beta} + \left(\frac{(50\sigma^2 + 8b\rho) \log n}{n}\right)^{1/2} + \frac{4\mathcal{L}}{n^{\alpha/2}}. \tag{1.5}$$

[5] $\mathbb{H}_{\alpha,\mathcal{L}}$ is the set of functions $W : [0,1]^2 \to \mathbb{R}$ satisfying $|W(x,y) - W(x',y')| \leqslant \mathcal{L}\big((x-x')^2 + (y-y')^2\big)^{\alpha/2}$ for all $x, y, x', y' \in [0,1]$.

We can also present specific instances for the different distributions discussed in table 1.1, specifically in the context of Lipschitz graphons. It should be noted that for piecewise constant graphons, the obtained results will be the same as those presented in table 1.1, with the inclusion of an additional approximation error term $\rho\left(\sqrt{\frac{K}{n}}+\sqrt{\frac{L}{m}}\right)^{1/2}$ as outlined in (1.3). Once again, we also provide an adaptive method for unknown K and L in the case of piecewise constant graphon.

Distr. of H_{ij}	Values (σ^2, b)	Condition (1.4)	Risk Bound (1.5)
Bernoulli(ρ)	$(\rho, 1/3)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 n \log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(nm)^{1/4}} + \dfrac{8\sqrt{\rho \log m}}{\sqrt{m}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
Binomial$(N, \rho)/N$	$(\rho/N, 1/3N)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 N n \log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Nnm)^{1/4}} + \dfrac{8\sqrt{\rho \log m}}{\sqrt{Nm}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
Poisson$(T\rho)/T$	$(\rho/T, 1/3T)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 T n \log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Tnm)^{1/4}} + \dfrac{8\sqrt{\rho \log m}}{\sqrt{Tm}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
sub-Gauss(σ^2)	$(\sigma^2, 0)$	$\sigma^2 \geqslant \dfrac{3\mathcal{L}^2 n \log^4(2n)}{25m^3}$	$\dfrac{11\sqrt{\mathcal{L}\sigma}}{(nm)^{1/4}} + \dfrac{8\sigma\sqrt{\log m}}{\sqrt{m}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$

Table 1.2: Upper bound for Lipschitz-continuous graphons and various distributions, under the non symmetric framework, with the additionnal assumption that $n \geqslant m$.

Relaxed independence assumption Now, let us assume that the matrix $\mathbf{H}$ is generated according to a re-scaled graphon W^*, with independent rows $\boldsymbol{H}_i$ that sum to one as considered in the model described in assumption 2, where

$$\mathbb{E}[\mathbf{H}|\boldsymbol{U}, \boldsymbol{V}] = \boldsymbol{\Theta}^* \quad \text{with} \quad \Theta^*_{ij} = \frac{W^*(U_i, V_j)}{\sum_{\ell=1}^m W^*(U_i, V_\ell)}\,.$$

In the context of the full independence, we already know that the graphon W^* is unidentifiable. In addition to that, in the present context, we can multiply W^* by a constant without changing the distribution of $\mathbf{H}$. To address this, we define a new equivalence class, where two graphons W and W' are considered equivalent if and only if they satisfy the relation

$$W = C_x W'(\tau_1 \otimes \tau_2)$$

where C_x is a constant that could depend on the first variable x and τ_1 and τ_2 are bijections that preserve the Lebesgue measure. It is evident that two such graphons will produce the same matrix $\mathbf{H}$. Moving forward, we assume that $W^* \in \mathscr{C}$, where

$$\mathscr{C} = \left\{W, I_W(x) = 1/m, \forall x \in [0,1]\right\} \quad \text{with} \quad I_W(x) = \int_0^1 W(x,y), dy.$$

Similar to before, the chosen distance within the class $\mathscr{C}$ to measure the quality of estimators is defined as

$$\begin{aligned}\delta(W', W) &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \left(\iint_{[0,1]^2} \left| W'(\tau_1(u), \tau_2(v)) - W(u,v) \right|^2, du, dv \right)^{\frac{1}{2}} \\ &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \|W' \circ (\tau_1 \otimes \tau_2) - W\|_{\mathbb{L}^2}.\end{aligned}$$

Estimating a graphon becomes a challenging task due to the intricate process of normalization within the class $\mathscr{C}$. In the subsequent theorem, we present an upper bound for the estimation of piecewise constant graphons, in the symmetric framework.

Theorem 4. *Let $\mathbf{H} \in [0,1]^{n\times n}$ be a $n \times n$ random matrix satisfying assumption 2 with some graphon $W^* : [0,1]^2 \to [0,\rho]$. Assume that each row of $\mathbf{H}$ sum to one, its covariance matrix $\mathbf{\Sigma}_i$ satisfies $\|\mathbf{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\mathbf{\Sigma}} \leqslant 1$ and its conditional mean matrix $\mathbf{\Theta}^*$ satisfies $\|\mathbf{\Theta}^*\|_\infty \leqslant \rho$. Assume that the graphon W^* is K-piecewise constant, meaning that for some integers $K \geqslant 2$ and for $0 = a_0 < \ldots < a_K = 1$, such that*

$$\Delta^{(K)} := \min_{k\in[K]} |a_k - a_{k-1}| \geqslant \frac{8\log(nK)}{n}$$

the function W^* is constant on each rectangle $[a_{k-1}, a_k[^2$. Then, the estimator $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}$ with $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0}[K]$ defined by (1.2) satisfies

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant \left(50\rho_{\mathbf{\Sigma}} + 6\rho\right)^{1/2} \left(\frac{3K^2}{n^2} + \frac{2\log K}{n}\right)^{1/2} + 3\rho\left(\frac{2K}{n}\right)^{1/4}$$

provided that $\psi_n(\Delta^{(K)}) = \frac{4\log(2e/\Delta^{(K)})}{n\Delta^{(K)}} \leqslant \rho_{\mathbf{\Sigma}}$ and $\bar{w} \leqslant \frac{1}{4}e^{0.045n}$ where $\bar{w} = \sum_{\ell=1}^{K} \frac{1}{w_k}$ and $w_k = a_k - a_{k-1}$.

The obtained result in this theorem is similar to the one presented in Theorem 3, with the additional requirement that $\bar{w}$ is not excessively large, ensuring that the size of the intervals $[a_k, a_{k+1}[$ is sufficiently large. For instance, if $n = 315$, the condition on $\bar{w}$ is satisfied as long as the minimum difference between consecutive a_k values is greater than or equal to 10^{-3}. This condition appears because of the aforementioned normalization in class $\mathscr{C}$.

1.4.3 Lower bound on the minimax risk

In this section, we establish the optimality of the least squares estimator $\widehat{W}^{\mathsf{LS}}$, under assumption 1 of full independence, by demonstrating its convergence rate in the worst-case scenario over the class $\mathcal{W}_\rho[K, L]$. This class consists of graphons W taht are constant over intervals I_k and J_ℓ, where $I_k = [a_k, a_{k+1})$ and $J_\ell = [b_\ell, b_{\ell+1})$ form a partition of $[0, 1)$.

We focus on proving the lower bound for the binomial model, but the techniques utilized in

the proof can be extended to the other models mentioned in the introduction. This establishes the optimality of the least squares estimator within this class.

Theorem 5. *Assume that conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the entries $H_{i,j}$ of the observed $n \times m$ matrix $\mathbf{H}$ are independent and drawn from the Binomial distribution with parameter $(N, W^*(U_i, V_j))$. There exist universal constants $c, C > 0$, such that for any $K, L > C$ satisfying $KL \geqslant L \log^2 L + K \log^2 K$ and for any $\rho > 0$,*

$$\inf_{\widehat{W}} \sup_{W^*} \mathbb{E}_{W^*} \left[\delta^2(\widehat{W}, W^*)\right]^{1/2} \geqslant c \left[\sqrt{\rho}\left(\frac{KL}{Nnm} \wedge \rho + \frac{1}{N\sqrt{nm}} \wedge \rho\right)^{1/2} + \rho\left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\right)^{1/2}\right],$$

where the inf is over all possible estimators $\widehat{W}$ and the sup is over all $W^* \in \mathcal{W}_\rho[K, L]$.

In this theorem, we could assume $NH_{i,j} \sim \mathcal{B}(N, W^*(U_i, V_j))$ – instead of assuming $H_{i,j} \sim \mathcal{B}(N, W^*(U_i, V_j))$) – and obtain the same result. For the balanced setting, when $n = m$, $K = L$, this lower bound has to be compared with (1.3) and appears to be rate optimal up to a $\log K$ factor. Figures 6 and 7 show the purple areas where the lower bound is of the order of the upper bound for various parameters of the model, that is our estimator is minimax optimal.

To be more accurate, in Figure 6, we fix the sparsity parameter ρ and choose cluster parameters K and L such that $K/n = L/m = \gamma$. The purple area in the figure represents pairs (n, m) where the lower bound obtained from Theorem 5 exceeds half of the upper bound given in (2.13). On the other hand, Figure 7 depicts the same criterion, but with fixed n and m, while varying ρ and γ. Notably, we observe that the least square estimator achieves optimality in many settings, even in highly asymmetric frameworks where for example, m is significantly larger than n.

1.4.4 Algorithm and numerical experiments

The least squares estimator introduced in equation (1.2) and discussed in the preceding sections is computationally intractable due to its combinatorial optimization nature. It is not feasible to compute this estimator in polynomial time. In this section, our objective is to present an algorithm that provides a computationally tractable approximation of $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}$. Although there is no guarantee that the algorithm always yields an estimator close to $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}$, it is expected to be the case in many scenarios.

The algorithm The proposed approximation can be seen as a variant of Lloyd's algorithm for k-means clustering [Llo82]. To describe it, let's recall that the least squares estimator is defined as a solution that minimizes the distance induced by the Frobenius norm between $\mathbf{H}$

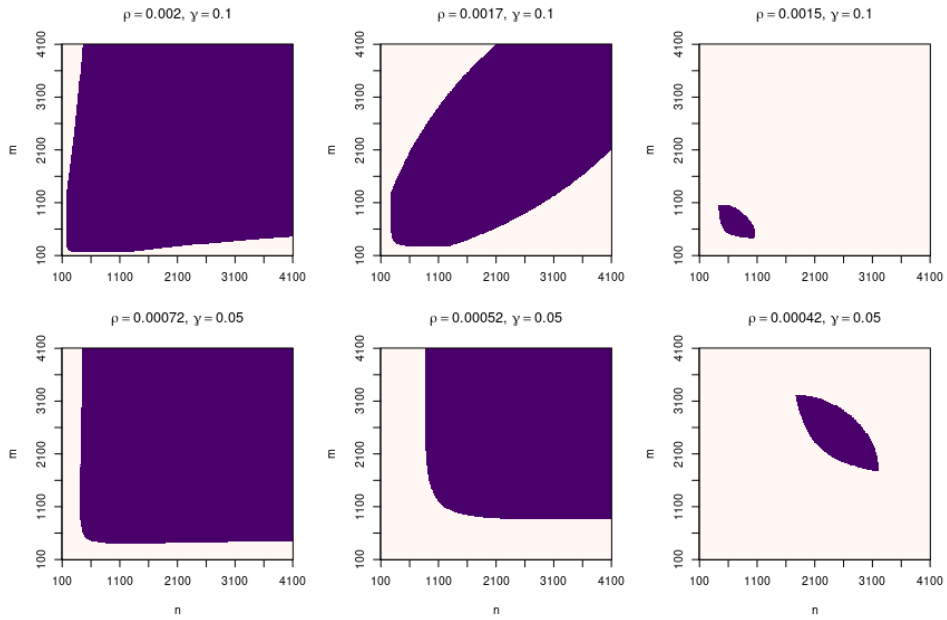


Figure 6: Illustration of the optimality of the least squares estimator for $N = 1$. The purple area corresponds to the values of n and m, for some fixed values of ρ and $\gamma = K/n = L/m$, for which the lower bound is within a constant factor of the upper bound. More precisely, when $\rho\gamma^2 \wedge \rho^2 + \rho(nm)^{-1/2} \wedge \rho^2 + 2\rho^2\sqrt{\gamma}$ is larger than half of $\rho\gamma^2 + (\rho \log K)/(3m) + (\rho \log L)/(3n) + 2\rho^2\sqrt{\gamma}$. We observe that unless ρ is very small, the upper bound established for the least-squares estimator is within a constant factor of the lower bound for all estimators for most values of n and m.

and a block constant matrix. This can be reformulated as follows

$$(\widehat{\mathbf{Q}}, \mathbf{Z}_1, \mathbf{Z}_2)^{\mathsf{LS}} \in \arg \min_{\substack{\mathbf{Q} \in \mathbb{R}^{K \times L} \\ \mathbf{Z}_1 \in \mathcal{Z}(n,K,n_0) \\ \mathbf{Z}_2 \in \mathcal{Z}(m,L,m_0)}} \left\| \mathbf{H} - \mathbf{Z}_1 \mathbf{Q} (\mathbf{Z}_2)^\top \right\|_{\mathsf{F}}^2. \tag{1.6}$$

where $\mathbf{Z}_1, \mathbf{Z}_2$ represent the block structure of the matrix[6] , that is the left and right clusters, and $\mathbf{Q}$ gives the values in the different blocks. It is interesting to note that when two of the three arguments $\mathbf{Q}$, $\mathbf{Z}_1$ or $\mathbf{Z}_2$ of the objective function are fixed, the minimization problem with respect to the remaining argument becomes computationally feasible. Therefore, we can employ the alternating minimization algorithm outlined below, which guarantees a decrease in the cost function $\mathcal{L}(\mathbf{Z}_1, \mathbf{Q}, \mathbf{Z}_2) = \|\mathbf{H} - \mathbf{Z}_1 \mathbf{Q} (\mathbf{Z}_2)^\top\|_{\mathsf{F}}^2$ at each iteration.

Initialization Procedure As shown in figure 8, the initial matrices chosen for algorithm 3 can significantly influence the final result. One approach to mitigate this issue is to run multiple instances of the algorithm in parallel, each with different initialization matrices randomly chosen. Among the resulting N estimators, the final estimator is selected as the one that minimizes the objective function $\mathcal{L}$.

[6] $\mathcal{Z}(n, K, n_0) = \left\{ \mathbf{Z} \in \{0,1\}^{n \times K} : \mathbf{Z}\mathbf{1}_K = \mathbf{1}_n \text{ and } \min_{k \in [K]} \mathbf{1}_n^\top \mathbf{Z}_{\bullet,k} \geqslant n_0 \right\}$ with $1_d = (1, \ldots, 1)^\top \in \mathbb{R}^d$.

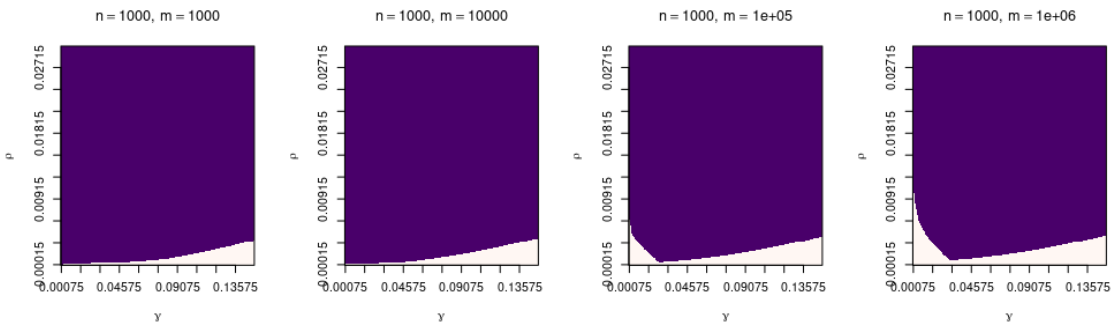


Figure 7: Illutration of the optimality of the least squares estimator. The purple area corresponds to the values of ρ and $\gamma = K/n = L/m$, for some fixed values of n and m, for which the lower bound is within a constant factor of the upper bound. We observe that unless ρ is very small, the upper bound and the lower bound for are of the same order.

Algorithm 3 Lloyd's algorithm of alternating minimization for approximating the LSE (1.6)

Require: $\mathbf{Z}_1$, $\mathbf{Z}_2$ the left and right cluster matrices with entries in $\{0, 1\}$, $\mathbf{H}$ the data matrix.
Ensure: $(\widetilde{\mathbf{Z}}_1, \mathbf{Q}, \widetilde{\mathbf{Z}}_2)$ local minimizer of $\mathcal{L}(\cdot, \cdot, \cdot)$.

Repeat :

1. Compute $\mathbf{Q} = (\mathbf{Z}_1^{\text{norm}})^\top \mathbf{H} \mathbf{Z}_2^{\text{norm}}$ where $\mathbf{Z}_1^{\text{norm}}$ is the matrix $\mathbf{Z}_1$ with normalized columns with respect to ℓ^1-norm (the number of 1 in the column), and similarly for $\mathbf{Z}_2^{\text{norm}}$.
2. Update $\mathbf{Z}_1$ that minimize $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}_2)$
3. Update $\mathbf{Z}_2$ that minimize $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z_1}, \mathbf{Q}, \mathbf{Z})$

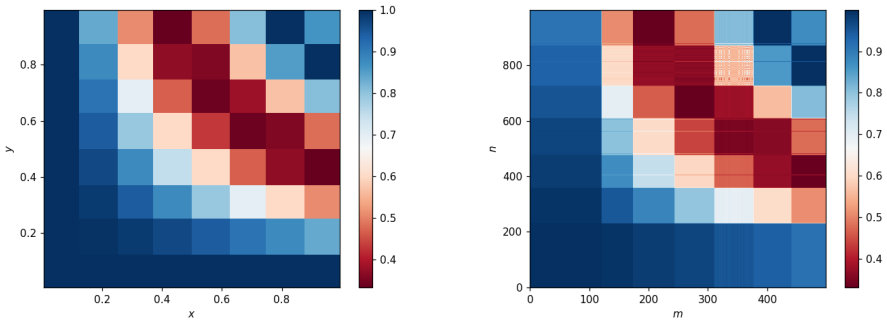
Another strategy, often used in conjunction with Lloyd's algorithm, is spectral initialization. In the case where the graphon is piecewise constant, the problem can be seen as a bi-stochastic block model for bipartite networks. One way to obtain initial values $(\mathbf{Z}_1, \mathbf{Z}_2)$ is through the spectral method proposed in [ZA19a]. This method involves computing the K-truncated singular value decomposition of a regularized version of the matrix $\mathbf{H}$. The K-truncated left singular vectors are then used as input for k-means clustering, resulting in an initialization for $\mathbf{Z}_1$. A similar procedure is applied to obtain the initialization for $\mathbf{Z}_2$. This spectral initialization approach can provide a good starting point for Lloyd's algorithm and improve the quality of the final estimator.

To assess the impact of the initialization procedure on the estimator given by algorithm 3, we conducted several runs using different cluster matrices. Specifically, we used matrices obtained through spectral clustering as mentioned earlier, matrices generated randomly once, and oracle clusters derived from the unobserved random variables. We then plotted the resulting estimators after rearranging the rows and columns based on permutations that ordered sequences of the unknown variables U_i and V_j. The results, depicted in Figure 8, clearly demonstrate that clustering initialization yields superior outcomes compared to a single random initialization.

Numerical experiments In this paragraph, we briefly present some numerical experiments to examine the behavior of the estimation error of the graphon $\widehat{W}^{\mathsf{LS}}$ and its relationship with various model parameters. We refer the reader to Section 2.6 for more precision about those numerical experiments. We begin by investigating the case of piecewise constant graphons and analyze the estimation error of the matrix $\mathbf{\Theta}^*$. We explore how this error varies with the parameter n for different values of (ρ, K, L), assuming $m = n/2$. Results are depicted in figure 9, where the values of the graphon W^* are randomly generated, and the error is drawing in the log-scale for different types of initialization already mentioned in the previous paragraph.

From the experimental results, it can be observed that the error of the "spectral" version consistently decreases as the value of n increases. Moreover, it converges to the oracle error at a faster rate when the sparsity parameter ρ is larger and when the ratios n/K and m/L are higher. This aligns with our intuition, as a higher ρ implies more links in the network, leading to more accurate estimation. Similarly, larger values of n/K and m/L also contribute to improved estimation accuracy. On the other hand, the "random" version of the algorithm exhibits a more erratic behavior. In most cases, its error surpasses that of the "spectral" version when n/K and m/L reach a certain threshold.

Additionally in figure 17, we present the estimation results for a Lipschitz-continuous graphon, where the parameters K and L are chosen as functions of n and m respectively, based on the recommendations provided by our theoretical findings. Interestingly, and somewhat surprisingly, the random initialization behaves as well as the spectral one. We do not have any explanation for this observation at this stage.



(a) Graphon W^*

(b) Alg 3 - s.c. init.

(c) Alg 3 - Rand. init. $(N = 1)$

(d) Alg 3 - Oracle init.

Figure 8: Illustration of the sensitivity of algorithm 3 to the initialization procedure. Plot 8a is the True Graphon W^*. Plot 8b is the obtained graphon after running algorithm 3 with spectral clustering matrices as initialization. Plot 8c is the obtained graphon with a random initilization of the algorithm. Plot 8d is the obtained graphon with the true (unknown) cluster matrices as initialization. The parameter chosen here are $(\rho, K, L, n, m) = (1, 8, 8, 1000, 500)$. We permuted the axes to plot the results. We can not do this in practice because we do not know the true permutations.

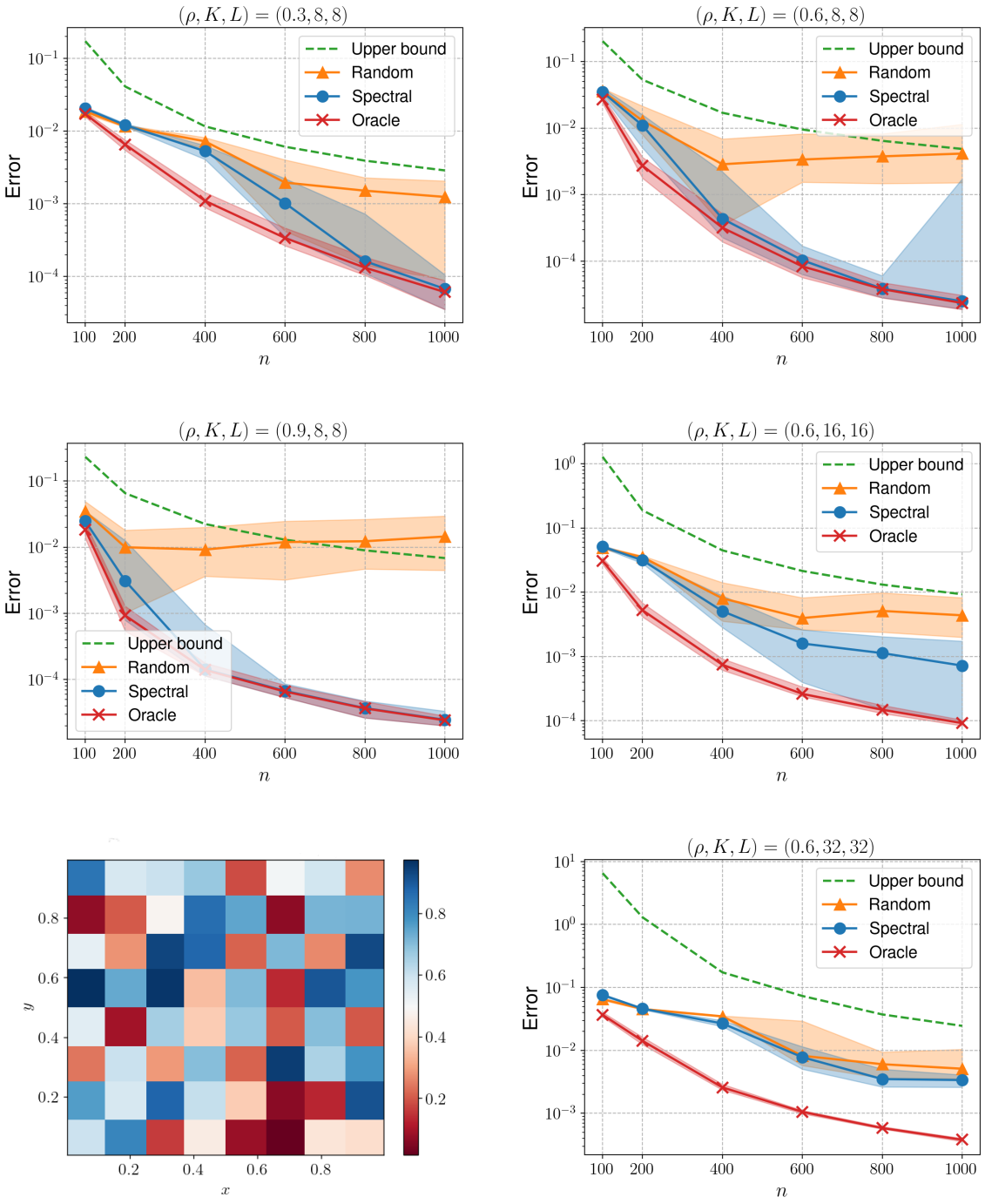
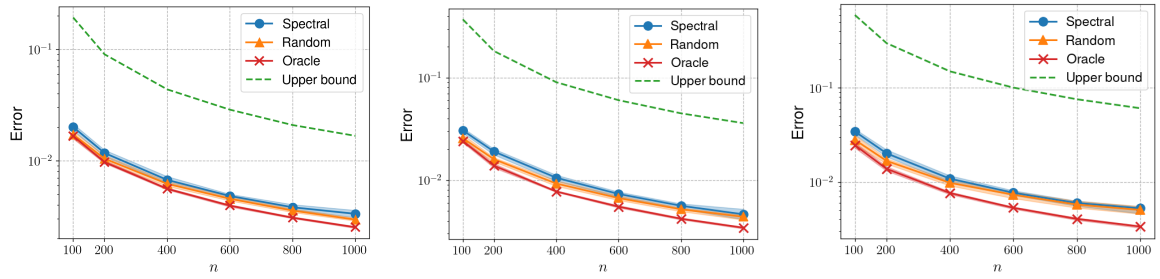
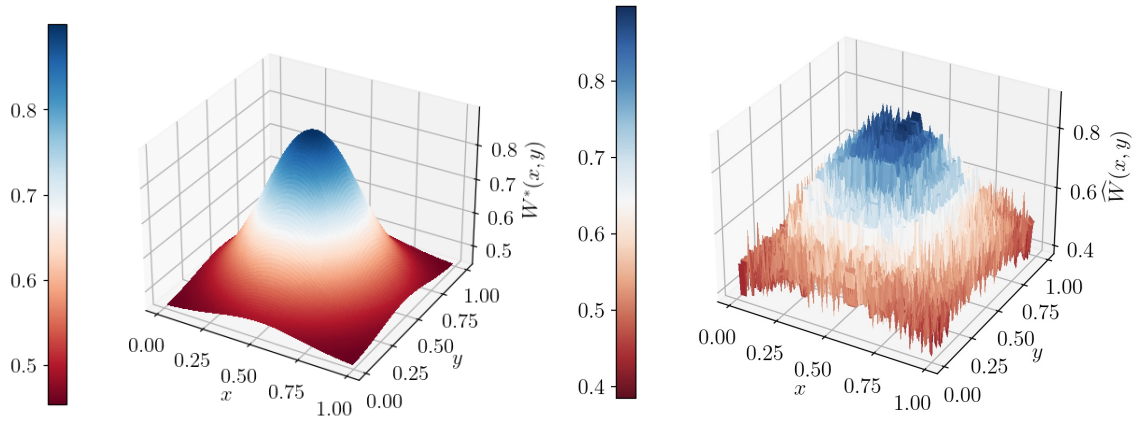


Figure 9: Evolution of the estimation error as a function of n, with $m = n/2$ for a piecewise constant random graphon for different values of (ρ, K, L).



(a) Error for $\rho = 0.3$ (b) Error for $\rho = 0.6$ (c) Error for $\rho = 0.9$



(d) True graphon W^* (e) Estimated graphon $\widehat{W}$

Figure 10: Evolution of the estimation error as a function of n, with $m = n/2$ for a Lipschitz graphon, for different values of ρ. The curves represent the error for various initializations of algorithm 3. The true graphon is represented in figure 10d and figure 10e is a representation of the rearranged estimated graphon with spectral initialisation. In practice, we can not rearranged the estimated graphon because it requires the knwoledge of the latent variables.

Chapter 2

Graphon estimation in bipartite graphs with observable edge labels and unobservable node labels

Abstract Many real-world data sets can be presented in the form of a matrix whose entries correspond to the interaction between two entities of different natures (number of times a web user visits a web page, a student's grade in a subject, a patient's rating of a doctor, etc.). We assume in this chapter that the mentioned interaction is determined by unobservable latent variables describing each entity. Our objective is to estimate the conditional expectation of the data matrix given the unobservable variables. This is presented as a problem of estimation of a bivariate function referred to as graphon. We study the cases of piecewise constant and Hölder-continuous graphons. We establish finite sample risk bounds for the least squares estimator and the exponentially weighted aggregate. These bounds highlight the dependence of the estimation error on the size of the data set, the maximum intensity of the interactions, and the level of noise. As the analyzed least-squares estimator is intractable, we propose an adaptation of Lloyd's alternating minimization algorithm to compute an approximation of the least-squares estimator. Finally, we present numerical experiments in order to illustrate the empirical performance of the graphon estimator on synthetic data sets.

2.1 Introduction

In this chapter [1], we consider the problem of estimating the conditional mean of a random matrix generated by a bivariate graphon and (unobserved) latent variables. More precisely, let n and m be two positive integers assumed to be large, and $\mathbf{H}$ be an $n \times m$ random matrix with real entries $H_{i,j}$. We assume that the distribution of this matrix $\mathbf{H}$ satisfies the following condition.

Assumption 3. There is a function $W^* : [0,1]^2 \to \mathbb{R}$, called the graphon, and two random vectors $\boldsymbol{U} = (U_1, \dots, U_n)$ and $\boldsymbol{V} = (V_1, \dots, V_m)$ such that

A 1.1 $U_1, \dots, U_n, V_1, \dots, V_m$ are independent and drawn from the uniform distribution $\mathcal{U}([0,1])$.

A 1.2 conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the entries $H_{i,j}$ are independent and $\mathbb{E}[H_{i,j}|\boldsymbol{U}, \boldsymbol{V}] = W^*(U_i, V_j)$.

The aforementioned setting corresponds to the practical situation in which there are n users and m items. Each user has an unobserved latent feature U and each item has an unobserved latent feature V. We observe the label H that characterizes the interaction between the user and the item. The function W^* corresponds to the mean value of the interaction for given values of the latent features.

[1] This chapter corresponds to a preprint paper accessible on arXiv [DMDK+23].

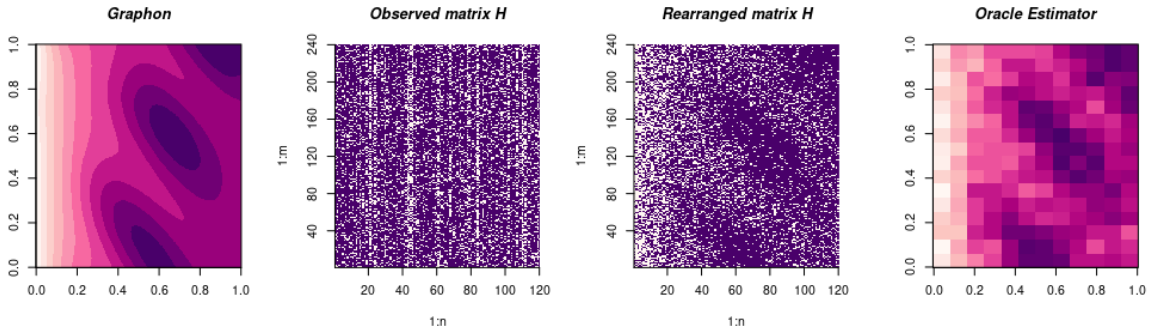


Figure 11: An illustration of the graphon problem. The leftmost graph represents the unknown graphon W^*. The second leftmost graph is the adjacency matrix observed in the graph where the links are made according to the Bernoulli model. The third graph is the adjacency matrix that would be obtained after a rearrangement of the rows and columns if we had access to the latent variables. The rightmost graph represents the histogram estimator obtained from the rearranged adjacency matrix. Our goal is to design an estimator which is nearly as good as the oracle, without having access to the latent variables.

The main examples to which the aforementioned setting as well as the results obtained in this chapter are applicable are the following:

1. The entries $H_{i,j}$ take the values 0 and 1 and correspond to the presence of an edge in the bipartite graph. In the example of customers and products, one might set $H_{i,j} = 1$ if and only if customer i has already bought the product j. It is important in this setting to take into consideration the case of large and sparse graphs, in which the probabilities of having an edge between two nodes are small for all pairs of nodes.
2. Each user is given N opportunities to interact with each item and the entries $H_{i,j}$ represent the empirical frequency of interaction. For instance, if users are players and items are games, each player plays each game $N = 10$ times, and $H_{i,j} = 3/10$ means that the player won the game 3 times out of 10. In this case $H_{i,j}$ are between 0 and 1. If the aforementioned N trials are independent, then the variance of $H_{i,j}$ is of order $1/N$. Thus, the specificity of this setting is that the variance of the noise is small. One can encompass this example in a more general setting of a sub-Gaussian distribution with a small variance proxy σ^2 corresponding to $1/N$.
3. $H_{i,j}$ is the average number of interactions between the user and the item over a time interval of length $T > 0$. More precisely, $H_{i,j} = N_{i,j}(T)/T$ where $N_{i,j}(T)$ is a Poisson random variable with intensity proportional to T. This parameter T could be large, thus allowing for a good estimation accuracy.

Our goal is to study the minimax risk of estimating W^* and to highlight its dependence on the important parameters of the problem. The sizes n and m of the matrix are among these parameters, but we will also be interested in the dependence on the smoothness of W^*, on the "sparsity" of interactions (denoted by ρ) and on the noise level (denoted by σ). These

parameters ρ and σ are positive real numbers such that

$$\|W^*\|_\infty = \sup_{u,v\in[0,1]} |W^*(u,v)| \leqslant \rho \qquad \text{and} \qquad \mathbf{Var}[H_{i,j}|U_i, V_j] \leqslant \sigma^2 \text{ a.s.}, \quad \forall i \in [n], \forall j \in [m].$$

The parameters ρ and σ may depend on n and m, but we choose to write ρ instead of $\rho_{n,m}$ and $\sigma_{n,m}$ for the sake of simplicity.

On the way to estimating the graphon W^*, an important intermediate step will consist in estimating the matrix $\mathbf{\Theta}^* = W^*(U_i, V_j)$. The estimation of this matrix is of interest on its own. We perform this task by solving the least squares problem over the set of constant-by-block matrices, with blocks generated by partitions of the sets of rows and the columns of the matrix $\mathbf{H}$. It will be further shown that the method of aggregation by exponential weights can be used to ensure adaptivity to the number of blocks. Under the condition that the graphon is piecewise constant or α-smooth in the sense of Hölder smoothness, we establish risk bounds for the graphon estimator derived from the estimator of $\mathbf{\Theta}^*$. These risk bounds are nonasymptotic, and shown to be rate optimal in the minimax sense for a broad range of regimes.

Measuring the quality of an estimator Figure 11 provides an illustration of the graphon estimation problem, in the case where $\mathbf{H}$ is the adjacency matrix of the bipartite graph, that is the entries of $\mathbf{H}$ are either 0 or 1. We see in this figure that the absence of knowledge of the latent variables has a strong impact on the recovery of the graphon. Indeed, the adjacency matrix $\mathbf{H}$ depicted in the second leftmost plot carries little information on W^*, as compared with the rearranged adjacency matrix displayed in the third plot. In fact, when $\boldsymbol{U}$ and $\boldsymbol{V}$ are unknown the graphon W^* is unidentifiable. Let us say that two graphons W and W' are equivalent, if there exist two bijections that preserve the Lebesgue measure $\tau_1 : [0,1] \to [0,1]$ and $\tau_2 : [0,1] \to [0,1]$ such that[2] $W = W' \circ (\tau_1 \otimes \tau_2)$. One can check that two matrices $\mathbf{H}$ generated by equivalent graphons W^* and $\bar{W}^*$ have the same distribution. This implies that one can at best estimate the equivalence class containing W^*. This is the reason underlying the (pseudo)-distance we use in this work when measuring the quality of an estimator $\widehat{W}$ of W^*, namely

$$\begin{aligned}\delta(\widehat{W}, W^*) &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \Big(\iint_{[0,1]^2} \big|\widehat{W}(\tau_1(u), \tau_2(v)) - W^*(u,v)\big|^2 du\, dv \Big)^{1/2} \\ &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \|\widehat{W} \circ (\tau_1 \otimes \tau_2) - W^*\|_{\mathbb{L}^2}\end{aligned}$$

where $\mathcal{M}$ is the set of all automorphisms $\tau : [0,1] \to [0,1]$ such that τ and τ^{-1} are measurable, and τ preserves the Lebesgue measure in the sense that $\lambda(\tau^{-1}(B)) = \lambda(B)$ for every Borel-set $B \subset [0,1]$. Two graphons W_1 and W_2 are called weakly isomorphic if $\delta(W_1, W_2) = 0$.

[2]We use notation $\tau_1 \otimes \tau_2$ for the function from $[0,1]^2$ to $[0,1]^2$ defined by $(\tau_1 \otimes \tau_2)(u,v) = (\tau_1(u), \tau_2(v))$.

Our contributions The main contributions of the present chapter are the following:

- We present a nonparametric framework based on bivariate graphon functions and unobservable latent variables that offer a flexible way of modeling random matrices and, in particular, adjacency matrices of random bipartite graphs.
- We establish finite sample risk bounds for the estimator minimizing the squared error over piecewise constant matrices with a given number of clusters, as well as for the exponentially weighted aggregate that combines the mentioned least-squares estimators. These results apply to adjacency matrices of bipartite graphs, whose entries are drawn from the Bernoulli distribution, but they are also valid for the binomial distribution, the scaled Poisson distribution and sub-Gaussian distributions.
- We present an adaptation of Lloyd's algorithm of alternating minimization (including a step of convex relaxation) to our setting, which allows us to obtain a computationally tractable approximation of the least squares estimator.
- In the case of matrices with entries having a Binomial conditional distribution given latent variables, we prove lower bounds on the worst-case risk, over the set of piecewise constant graphons, for any graphon estimator. These lower bounds, in the vast majority of cases, are of the same order as the upper bounds obtained for the least-squares estimator.

Prior work Statistical analysis of matrix and network data, based on block models similar to those considered in this work, is an active area of research since at least two decades. The stochastic block model, introduced by [HLL83], is perhaps one of the most studied latent structures for network data, see also [NS01] and [GN03] for early references. Community detection, which is the problem of detecting the underlying block structure, has been the focus of much research effort, as evidenced by studies such as [ZLZ12, CRV15, LR15, Lei16, ZZ16, WB17, CLX18, XJL20] and literature reviews such as [GMZZ17, Abb18]. It should be noted that the majority of these studies focused on unipartite graphs with binary or discrete edge-labels, and their optimality was mostly related to identifying the smallest separation rate between the parameters of the communities that enables their consistent recovery. Similar problems for bipartite block models have been investigated in [FPV15, FP16b, Neu18, ZA19b, ZA20, CLC+21, NST22]. Several studies, such as the one conducted by [CDP12], have delved into the estimation of parameters in a stochastic block model using accessible techniques like variational inference. Notably, [GK21] presents a groundbreaking contribution by introducing the first minimax-optimal and tractable estimator for parameter estimation in the stochastic block model, especially when dealing with missing links. Additionally, [LM19] underscores the efficacy of variational methods in approximating the maximum likelihood estimator for dynamic stochastic block models.

In contrast with the aforementioned papers, the focus here is on optimality in terms of

the estimation error for a model that encompasses bipartite graphs with real-valued edge labels. Consistency of graphon estimators has been studied in [ACC13, WO13, OW14a]. In the same problem, minimax-rate-optimality of the least squares estimator has been established in [GLZ15, GLMZ16, KTV17, KV19], see also the survey article [GM21]. To better present our contributions within the current state-of-the-art, it is useful to provide a brief overview of the contents of these papers. [GLZ15] considered the case of binary observations $H_{i,j}$, focusing on the dense case $\rho \asymp 1$, and obtained minimax rates of estimation over the classes of piecewise constant and Hölder continuous graphons. [GLMZ16] extended these results to matrices $\mathbf{H}$ with sub-Gaussian entries, some of which might be missing completely at random. However, as discussed in Subsection 2.3.1, their results are sub-optimal in some cases w.r.t. the noise variance, for instance when edge-labels are drawn from the binomial distribution. In the case of a unipartite graph with binary edge-labels, [KTV17] established the minimax-optimal rates of estimation for sparsely connected graphs, where $\rho \ll 1$. While [GLZ15, GLMZ16] measured the estimation error using the normalized Frobenius norm of the difference between the estimated matrix and the true one, [KTV17] additionally considered the $\mathbb{L}_2$-distance between the equivalence classes of graphons. In [KV19], minimax optimal rates of graphon estimation in the cut distance have been established for unipartite graphs with binary observations.

The problems studied in this work have connections with some recent work in econometrics. Indeed, as a consequence of [Ald81, Theorem 1.4], if the matrix $\mathbf{H}$ is row and column exchangeable, then there is a function $g^* : [0,1]^4 \to \mathbb{R}$ and independent random variables α, $\{U_i\}$, $\{V_j\}$, $\{\xi_{ij}\}$ uniformly distributed in $[0,1]$ such that the random matrices $\mathbf{H}$ and $\big(g^*(\alpha, U_i, V_j, \xi_{ij}); i \in [n], j \in [m]\big)$ have the same distribution. The problem under consideration in this chapter is equivalent to estimating the random function $W^*(u,v) = \int_0^1 g^*(\alpha, u, v, z)\, dz$ from the observations $g^*(\alpha, U_i, V_j, \xi_{ij}); i \in [n], j \in [m]$, without assuming any parametric form of the function g^*. If in addition to $H_{i,j}$, we are also given a feature vector $\boldsymbol{X}_{i,j}$, for every pair of nodes (i,j) then the extended model defined by $g^*(\boldsymbol{X}_{i,j}, \alpha, U_i, V_j, \xi_{i,j})$ can be considered. This approach is adopted, for instance, in [Gra17], where the specific parametric form $g^*(\boldsymbol{x}, \alpha, u, v, z) = \mathbb{1}\big(\boldsymbol{x}^\top \boldsymbol{\beta} + u + v + \log(z/(1-z))\big)$ is considered. In such a parametric context, the parameter of interest is the vector $\boldsymbol{\beta}$. In [Gra20], it is assumed that the regression function $\int_{[0,1]^3} g^*(\boldsymbol{x}, \alpha, u, v, z)\, du\, dv\, dz$ has a parametric form $\exp\{a + \boldsymbol{x}^\top \boldsymbol{\beta}\}/(1 + \exp\{a + \boldsymbol{x}^\top \boldsymbol{\beta}\})$ and the problem of estimating the vector $(a, \boldsymbol{\beta})$ is studied. Asymptotic results (law of large numbers and central limit theorem) for exchangeable arrays have been proven in [DDG21b].

Notation For an integer $n \geqslant 1$, we set $[n] = \{1, \dots, n\}$. In mathematical formulae, we use bold capitals for matrices and bold italic letters for vectors. The integer part of a real number x is denoted by $\lfloor x \rfloor$, whereas the minimum and the maximum of two real number x, y are denoted by $x \wedge y$ and $x \vee y$, respectively. For two $n \times m$ matrices $\mathbf{B}$ and $\bar{\mathbf{B}}$, the inner product is defined

as

$$\langle \mathbf{B}, \bar{\mathbf{B}} \rangle = \mathrm{tr}(\mathbf{B}\bar{\mathbf{B}}^\top) = \sum_{i=1}^{n}\sum_{j=1}^{m} B_{ij}\bar{B}_{ij},$$

and we denote by $\|\mathbf{B}\|_{\mathsf{F}} = \sqrt{\langle \mathbf{B}, \mathbf{B} \rangle}$ the Frobenius norm of the matrix $\mathbf{B}$. The sup-norm of $\mathbf{B}$ denoted by $\|\mathbf{B}\|_\infty$ is defined as the largest in absolute value entry of $\mathbf{B}$. We write $\boldsymbol{B}_{i,\bullet}$ and $\boldsymbol{B}_{\bullet,j}$ for the ith row and the jth column of $\mathbf{B}$, respectively. The length of an interval $I \subset \mathbb{R}$ is denoted by $|I|$. For $n \in \mathbb{N}$, $\mathbf{1}_n$ is the n-vector with all its entries equal to one.

2.2 Estimators of the mean matrix and the graphon

In this section, we define the estimators of the mean matrix $\boldsymbol{\Theta}^* = \mathbf{E}[\mathbf{H}|\boldsymbol{U}, \boldsymbol{V}]$ and of the graphon W^* that are investigated in this chapter. We focus here on mathematical definitions only; computational and algorithmic properties of these estimators and their tractable approximations are deferred to Section 2.4.

2.2.1 Least squares estimator of $\boldsymbol{\Theta}^*$

Let us start by introducing some notation. For positive integers n_0, n, K satisfying $Kn_0 \leqslant n$ and $K \geqslant 2$, we define the set

$$\mathcal{Z}(n, K, n_0) = \Big\{ \mathbf{Z} \in \{0,1\}^{n\times K} : \mathbf{Z}\mathbf{1}_K = \mathbf{1}_n \text{ and } \min_{k\in[K]} \mathbf{1}_n^\top \boldsymbol{Z}_{\bullet,k} \geqslant n_0 \Big\}. \tag{2.1}$$

The elements of this set can be seen as assignment matrices: each one of the n users is assigned to one (and only one) of the K "communities", and we have the condition that each community has at least n_0 "members". Similarly, we will repeatedly use the set $\mathcal{Z}(m, L, m_0)$ of the assignment matrices corresponding to the items. Since the n rows of $\mathbf{H}$ correspond to the users and the m columns of $\mathbf{H}$ correspond to items, the elements of $\mathcal{Z}(n, K, n_0)$ will be denoted by $\mathbf{Z}^{\mathrm{user}}$ whereas the elements of $\mathcal{Z}(m, L, m_0)$ will be denoted by $\mathbf{Z}^{\mathrm{item}}$. Matrices $\mathbf{Z}^{\mathrm{user}}$ and $\mathbf{Z}^{\mathrm{item}}$ correspond to a biclustering: the clusters of users are specified by the matrix $\mathbf{Z}^{\mathrm{user}}$; in the same way, $\mathbf{Z}^{\mathrm{item}}$ encodes the clusters of items.

Given the observed adjacency matrix $\mathbf{H}$, the least squares estimator is defined by

$$(\widehat{\mathbf{Q}}, \widehat{\mathbf{Z}}^{\mathrm{user}}, \widehat{\mathbf{Z}}^{\mathrm{item}})^{\mathsf{LS}} \in \arg \min_{\substack{\mathbf{Q}\in\mathbb{R}^{K\times L} \\ \mathbf{Z}^{\mathrm{user}}\in\mathcal{Z}(n,K,n_0) \\ \mathbf{Z}^{\mathrm{item}}\in\mathcal{Z}(m,L,m_0)}} \big\|\mathbf{H} - \mathbf{Z}^{\mathrm{user}}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top\big\|_{\mathsf{F}}^2. \tag{2.2}$$

Here, $\mathbf{Z}^{\mathrm{user}}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top$ is a $n \times m$ constant-by-block matrix. The idea is thus to find the constant-by-block matrix that is the closest to $\mathbf{H}$ in the metric induced by the Frobenius norm, where the

blocs are given by the matrices $\mathbf{Z}^{\mathrm{user}}$ and $\mathbf{Z}^{\mathrm{item}}$, and the sizes of blocks are at least $n_0 \times m_0$.

These estimators computed by (2.2) lead to the constant-by-block least squares estimator of $\mathbf{\Theta}^*$ defined by $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{Z}}^{\mathrm{user}}\widehat{\mathbf{Q}}(\widehat{\mathbf{Z}}^{\mathrm{item}})^\top$. One can write $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ in the following alternative way. Let us consider the class of constant-by-block matrices

$$\mathcal{T} = \mathcal{T}^{K,L}_{n_0,m_0} = \left\{\mathbf{\Theta} = \mathbf{Z}^{\mathrm{user}}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top : \ (\mathbf{Q}, \mathbf{Z}^{\mathrm{user}}, \mathbf{Z}^{\mathrm{item}}) \in \mathbb{R}^{K\times L} \times \mathcal{Z}_{n,K,n_0} \times \mathcal{Z}_{m,L,m_0}\right\}. \quad (2.3)$$

The least squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ is a solution to

$$\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K,L] \in \arg\min_{\mathbf{\Theta}\in\mathcal{T}^{K,L}_{n_0,m_0}} \|\mathbf{H} - \mathbf{\Theta}\|_{\mathsf{F}}^2. \quad (2.4)$$

Our first results, reported in the next section, provide non asymptotic upper bounds on the risk of the estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$.

2.2.2 Aggregation by exponential weights

The least squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ is constant on KL blocks. The number of these blocks, chosen beforehand, is a hyperparameter of the method. If the true matrix $\mathbf{\Theta}^*$ is far from being blockwise constant on KL blocks, then the quality of estimation by $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K,L]$ might be poor because of the presence of a large bias. One can reduce this bias by computing the least squares estimator for several values of K, L (but also n_0 and m_0) and then by aggregating these estimators.

To this end, we consider here the extended framework in which two independent copies $\mathbf{H}$ and $\mathbf{H}'$ are observed, both satisfying Assumption A 1.2 (with exactly the same $\boldsymbol{U}$ and $\boldsymbol{V}$). The matrix $\mathbf{H}$ is used to construct estimators, whereas $\mathbf{H}'$ is used to define "weights" which are used for computing the exponentially weighted aggregate (EWA). More precisely, we denote by $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}_1, \ldots, \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_s$ least -squares estimators computed by solving (2.4) for s different values of (K, L, n_0, m_0). We define

$$\widehat{\mathbf{\Theta}}^{\mathsf{EWA}} = \sum_{\ell=1}^{s} w_\ell \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_\ell, \quad \text{with} \quad w_\ell = \frac{\exp\{\|\mathbf{H}' - \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_\ell\|_{\mathsf{F}}^2/\beta\}}{\sum_{r=1}^{s}\exp\{\|\mathbf{H}' - \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_r\|_{\mathsf{F}}^2/\beta\}} \quad (2.5)$$

where $\beta > 0$ is a parameter often referred to as the temperature. Since $\mathbf{H}'$ and $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}_\ell$ have been computed on two independent data matrices, they are independent; this will play an important role in the proofs. The choice of β depends on the nature of the observations and, more precisely, on the distribution of the noise $\boldsymbol{\xi}$, see the next section for more details.

2.2.3 Adaptations in the case of missing values

The estimators of $\mathbf{\Theta}^*$ presented in previous paragraphs use all the entries of the matrix $\mathbf{H}$. However, these estimators, as well as the mathematical results stated in the next sections, are easy to adapt to the case of missing observations. More precisely, assume that we observe some iid random variables $M_{i,j}$ taking values 0 and 1 such that the value $H_{i,j}$ is revealed to the statistician if and only if $M_{i,j} = 1$. Denoting by $\mathsf{p} = \mathbf{P}(M_{i,j} = 1)$ and assuming that $M_{i,j}$ is independent of $(H_{i,j}, U_i, V_j)$ (this case is commonly referred to as missing completely at random), we can define the adjusted observation matrix $\widetilde{\mathbf{H}}$ by its entries $\widetilde{H}_{i,j} = H_{i,j}M_{i,j}/\mathsf{p}$, for $i \in [n]$ and $j \in [m]$. Here, we assume that p is known. If this is not the case, we can easily estimate it using the empirical mean of the random variables $M_{i,j}$.

Then, to define the least-squares estimator of $\mathbf{\Theta}^*$, it suffices to replace $\mathbf{H}$ with $\widetilde{\mathbf{H}}$ in (2.2). Similar modifications can be made for defining the exponentially weighted aggregate. Note that this strategy has been already successfully applied in [GLMZ16]. The entries of matrix $\widetilde{\mathbf{H}}$ are all observable, the conditional on $(\boldsymbol{U}, \boldsymbol{V})$ expectation of $\widetilde{\mathbf{H}}$ is still $\mathbf{\Theta}^*$, and $\mathbf{Var}[\widetilde{H}_{i,j}|U_i, V_j] \leqslant \big(\sigma^2 + \rho^2(1-\mathsf{p})\big)/\mathsf{p}$, where σ^2 is an upper bound on the conditional variance $\mathbf{Var}[H_{i,j}|U_i, V_j]$ and ρ^2 is an upper bound on $(\Theta^*_{i,j})^2$.

2.2.4 Estimating the graphon

Having estimated the matrix $\mathbf{\Theta}^*$, the focus shifts to developing an estimator for the graphon W^*. To this end, for any $n \times m$ matrix $\mathbf{\Theta}$, we define its associated graphon $W_{\mathbf{\Theta}} : [0,1]^2 \to [0,1]$ as a constant function on each rectangle $I_i \times J_j = [(i-1)/n, i/n[\times[(j-1)/m, j/m[$, for $(i,j) \in [n] \times [m]$, given by

$$W_{\mathbf{\Theta}}(u,v) = \Theta_{i,j}, \quad \text{for all} \quad (u,v) \in I_i \times J_j. \tag{2.6}$$

The rationale behind this definition is the following: when n and m are large, the order statistics $U_{(i)}$ and $V_{(j)}$ lie with high probability in the intervals I_i and J_j. Therefore, the matrix $\widetilde{\mathbf{\Theta}}$ defined by $\widetilde{\Theta}_{i,j} = W_{\mathbf{\Theta}}(U_i, V_j)$ coincides, up to a permutation of rows and a permutation of columns, with $\mathbf{\Theta}^*$. This means that the matrices generated by W^* and $W_{\mathbf{\Theta}^*}$ are equivalent. Hence, one can expect that these two graphons are close.

In addition, for an estimated graphon defined by (2.6), the estimation error can be easily related to the error, measured by the Frobenius norm, in estimating matrix $\mathbf{\Theta}^*$. Indeed, one easily checks that $\|W_{\widehat{\mathbf{\Theta}}} - W_{\mathbf{\Theta}^*}\|_{\mathbb{L}_2} = \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}/\sqrt{nm}$, which leads to

$$\delta(W_{\widehat{\mathbf{\Theta}}}, W^*) \leqslant \|W_{\widehat{\mathbf{\Theta}}} - W_{\mathbf{\Theta}^*}\|_{\mathbb{L}_2} + \delta(W_{\mathbf{\Theta}^*}, W^*) \leqslant \frac{\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + \delta(W_{\mathbf{\Theta}^*}, W^*). \tag{2.7}$$

We will use this inequality both for $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ and $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}^{\mathsf{EWA}}$. To ease notation, we often

write $\widehat{W}^{\mathrm{LS}}$ and $\widehat{W}^{\mathrm{EWA}}$ instead of $W_{\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}}$ and $W_{\widehat{\boldsymbol{\Theta}}^{\mathrm{EWA}}}$, respectively. The decomposition provided by (2.7) splits the graphon estimation error into two components: the error of estimating the conditional mean matrix $\boldsymbol{\Theta}^*$ and the bias of approximating W^* by the piecewise constant function $W_{\boldsymbol{\Theta}^*}$. The former is the only term that depends on the estimation routine and on the probabilistic assumptions on the noise; it will be analyzed in the next section under various such assumptions. The latter depends only on the "smoothness properties" of the graphon. The next result allows us to evaluate this term.

Proposition 2. *Let $\Theta^*_{ij} = W^*(U_i, V_j)$ for $i \in [n]$ and $j \in [m]$, where $W^* : [0,1]^2 \to [A,B]$ for some A, B such that $-\infty \leqslant A < B \leqslant +\infty$.*

1. *(Piecewise constant graphon) More precisely, for $0 = a_0 < \ldots < a_K = 1$ and $0 = b_0 < \ldots < b_L = 1$, the function W^* is constant on each rectangle $[a_k, a_{k+1}[\times[b_\ell, b_{\ell+1}[$. If we define $W_{\boldsymbol{\Theta}^*} : [0,1]^2 \to [A,B]$ by $W_{\boldsymbol{\Theta}^*}(u,v) = \Theta^*_{i,j}$ for all $(u,v) \in [(i-1)/n, i/n[\times[(j-1)/n, j/n[$, then*

$$\mathbb{E}[\delta(W_{\boldsymbol{\Theta}^*}, W^*)^2]^{1/2} \leqslant \frac{(B-A)}{\sqrt{2}}\left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\right)^{1/2}.$$

2. *(Hölder continuous graphon) If the graphon W^* is α-smooth, that is W^* is in the Hölder class[3] $\mathbb{H}_{\alpha,\mathcal{L}}$, for some $\alpha \in (0,1]$ and $\mathcal{L} > 0$, then*

$$\mathbb{E}[\delta(W_{\boldsymbol{\Theta}^*}, W^*)^2]^{1/2} \leqslant \frac{2\mathcal{L}}{n^{\alpha/2}} + \frac{2\mathcal{L}}{m^{\alpha/2}}.$$

The proof of this result, postponed until Subsection 2.7.2, follows essentially the steps of [KTV17, Proposition 3.2] that deals with symmetric functions only. As a minor remark, the constants in our results are smaller than those available in the literature.

2.3 Finite sample risk bounds

We have introduced in Section 2.2 the least-squares estimator and the exponentially weighted aggregate for estimating the matrix $\boldsymbol{\Theta}^*$, as well as their associated graphon estimators. We provide in this section upper bounds for the risks of these estimators. The main purpose of these bounds is to highlight the behavior of the estimators when n, m are large and ρ, σ are small (σ denoting the noise magnitude).

[3] $\mathbb{H}_{\alpha,\mathcal{L}}$ is the set of functions $W : [0,1]^2 \to \mathbb{R}$ satisfying $|W(x,y) - W(x',y')| \leqslant \mathcal{L}\big((x-x')^2 + (y-y')^2\big)^{\alpha/2}$ for all $x, y, x', y' \in [0,1]$.

2.3.1 Risk bounds for the least-squares estimator $\widehat{\Theta}^{\mathsf{LS}}$

We start by stating risk bounds for estimators of $\mathbf{\Theta}^*$. To this end, without loss of generality, both in the statements and in the proofs, we treat $\mathbf{\Theta}^*$ as a deterministic matrix; this is why we require $H_{i,j}$ to be independent instead or requiring conditional independence given $\boldsymbol{U}, \boldsymbol{V}$.

Theorem 6. *Let n, m, n_0, m_0, L, K be positive integers such that $L \geqslant 2$, $K \geqslant 2$, $3 \leqslant n_0 \leqslant n$ and $3 \leqslant m_0 \leqslant m$. Let $\mathbf{H}$ be an $n \times m$ random matrix with independent entries satisfying $\mathbb{E}[H_{ij}] \in [0, \rho]$ for every $i \in [n]$, $j \in [m]$ and some $\rho > 0$. In addition, assume that the random variables $(H_{ij} - \mathbb{E}[H_{ij}])$ satisfy the (σ^2, b)-Bernstein condition. Then, the least squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ of the mean matrix $\mathbf{\Theta}^* = \mathbb{E}[\mathbf{H}]$, defined by (2.4), satisfies the exact oracle inequality*

$$\frac{\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \inf_{\mathbf{\Theta} \in \mathcal{T}^{K,L}_{n_0,m_0}} \frac{\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + \big(25\sigma^2 + 4b\rho\big)^{1/2} r_{n,m}(K, L),$$

with $r_{n,m}(K, L)$ given by

$$r_{n,m}(K, L) = \Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} \tag{2.8}$$

provided that $\psi_{n,m}(n_0, m_0) := \frac{3}{m_0} \log(en/n_0) + \frac{3}{n_0} \log(em/m_0) \leqslant (\sigma/b)^2$.

Lower bounds on the minimax risk of all possible estimators, showing that the risk bound in Theorem 8 is rate optimal under various regimes, will be stated in Section 2.5. Let us mention here the fact that in the particular case $b = 0$ corresponding to a sub-Gaussian distribution, the condition $\mathbf{E}[H_{i,j}] \in [0, \rho]$ can be removed and the claim of the theorem remains true. This causes no problem since ρ appears in the upper bound through the product $b \times \rho$ only.

Theorem 8 being stated for general distributions, it is helpful to see its consequences in the cases of common distributions of $H_{i,j}$ mentioned in the introduction.

Corollary 1. We assume that the conditions on n, m, n_0, m_0, L, K required in Theorem 6 hold.

1. If H_{ij}'s are independent Bernoulli—or any other distribution with support $[0, 1]$— random variables with mean $\Theta^*_{i,j} \leqslant \rho$, then they satisfy the $(\rho, 1/3)$-Bernstein condition and, therefore,

$$\frac{\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \inf_{\mathbf{\Theta} \in \mathcal{T}^{K,L}_{n_0,m_0}} \frac{\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + 9\sqrt{\rho}\Big(\frac{KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} \tag{2.9}$$

   provided that $\frac{1}{m_0} \log(en/n_0) + \frac{1}{n_0} \log(em/m_0) \leqslant 3\rho$.

2. If for some $N \in \mathbb{N}$, (NH_{ij})'s are independent binomial random variables with parameters $(N, \Theta^*_{i,j})$ such that $\Theta^*_{i,j} \leqslant \rho$, then $H_{i,j}$'s satisfy the $(\rho/N, 1/3N)$-Bernstein condition and,

therefore,

$$\frac{\mathbb{E}\big[\|\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \inf_{\boldsymbol{\Theta} \in \mathcal{T}^{K,L}_{n_0,m_0}} \frac{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + 9\sqrt{\rho}\left(\frac{KL}{Nnm} + \frac{\log K}{Nm} + \frac{\log L}{Nn}\right)^{1/2}$$

provided that $\frac{1}{m_0}\log(en/n_0) + \frac{1}{n_0}\log(em/m_0) \leqslant 3N\rho$.

3. If H_{ij}'s are independent sub-Gaussian random variables with means $\Theta^*_{i,j}$ and variance proxies $\tau^2_{i,j} \leqslant \sigma^2$, then they satisfy the $(\sigma^2, 0)$-Bernstein condition and, therefore,

$$\frac{\mathbb{E}\big[\|\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \inf_{\boldsymbol{\Theta} \in \mathcal{T}^{K,L}_{n_0,m_0}} \frac{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + 5\sigma\left(\frac{KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\right)^{1/2}.$$

4. If for some $T > 0$, (TH_{ij})'s are independent Poisson random variables with parameters $T\Theta^*_{i,j} \leqslant T\rho$, then $H_{i,j}$ satisfy the $(\rho/T, 1/3T)$-Bernstein condition and, therefore,

$$\frac{\mathbb{E}\big[\|\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \inf_{\boldsymbol{\Theta} \in \mathcal{T}^{K,L}_{n_0,m_0}} \frac{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + 9\sqrt{\rho}\left(\frac{KL}{Tnm} + \frac{\log K}{Tm} + \frac{\log L}{Tn}\right)^{1/2}$$

provided that $\frac{1}{m_0}\log(en/n_0) + \frac{1}{n_0}\log(em/m_0) \leqslant 3T\rho$.

The expression of the remainder term appearing in these risk bounds can be seen as

$$\text{noise magnitude} \times \frac{\text{size of the parameter space}}{\text{sample size}}.$$

Indeed, $KL + n\log K + m\log L$ is the order of magnitude of the logarithm of the covering number of $\mathcal{T}^{K,L}_{n_0,m_0}$, a common measure of the complexity of the parameter space. In addition to being instructive, this interpretation explains why this upper bound is optimal up to a multiplicative constant under some mild conditions.

Note also that the least-squares estimator for which the risk bounds above are established does not require the knowledge of ρ, σ, and b. This explains the presence of a condition on ρ requiring it to be not too small. For small values of ρ, our proofs may still be used to get a risk bound for the least-squares estimator. For instance, in the Bernoulli model, when $\rho \leqslant \frac{1}{m_0}\log(en/n_0) + \frac{1}{n_0}\log(em/m_0)$, the remainder term is the same as in (2.9) with ρ replaced by $\frac{1}{m_0}\log(en/n_0) + \frac{1}{n_0}\log(em/m_0)$. However, for such a small value of ρ smaller risk bounds can be obtained either for the estimator that outputs a matrix with all zero entries, or for the constrained least-squares estimator with the constraint $\|\boldsymbol{\Theta}\|_\infty \leqslant \rho$ (see [GLMZ16, KTV17] for results with this flavor).

Remark 1. As mentioned in the introduction, an upper bound similar to the one of Theorem 6 has been established in [GLMZ16], under the condition that $H_{i,j} - \mathbb{E}[H_{i,j}]$ are σ-sub-Gaussian. The remainder term obtained therein is of the order $(\sigma + \rho)r_{n,m}(K, L)$. Since, σ being sub-Gaussian is equivalent to the $(\sigma^2, 0)$-Bernstein condition, our theorem applies to the same

setting and yields a smaller remainder term, $\sigma r_{n,m}(K,L)$ (which is independent of ρ). Of particular interest is the case where the entries are scaled binomial random variables, as in the second claim of the last corollary. In this scenario, the risk bound of [GLMZ16] includes a remainder term of order $r_{n,m}(K,L)$ since the sub-Gaussian norm of the averages of independent Bernoulli random variables is of constant order. Interestingly, our upper bound is substantially tighter since its remainder term includes a deflation factor of $\sqrt{\rho/N}$. In Section 2.5, we demonstrate that this upper bound is tight, at least when n and m are of the same order of magnitude.

2.3.2 Risk bounds for the EWA $\widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}}$

In a framework pertaining to denoising the observed signal, [LB06] were the first to establish sharp bounds. They did so for Gaussian noise only. Extensions to more general noise distributions were developed in [DT07, DT08, DT12, Dal20, Dal22]. The results of this section are consequences of those presented in [Dal22], which are, to our knowledge, the only results in the literature applicable to models with asymmetric noise (as is the case in the Bernoulli and the binomial models).

Theorem 7. *Let $\mathbf{H}$ an $n \times m$ matrix with independent entries and let $\boldsymbol{\Theta}^* = \mathbf{E}[\mathbf{H}]$. Let $\mathbf{H}'$ be an independent copy of $\boldsymbol{H}$. Let $\mathfrak{P}$ be a set of quadruplets $(K, L, n_0, m_0) =: \mathfrak{p}$ and let $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}[\mathfrak{p}]$ be the least squares estimator defined by (2.4). Let $\widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}}[\mathfrak{P}]$ be the exponentially weighted aggregate (2.5) applied to estimators $\{\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}[\mathfrak{p}] : \mathfrak{p} \in \mathfrak{P}\}$ with some temperature parameter $\beta > 0$. Let $r_{n,m}(K,L)$ be as in (2.8).*

1. *(Bernoulli/binomial model) Assume that for some $N \in \mathbb{N}$, $NH_{i,j} \sim \mathsf{binomial}(N, \Theta^*_{i,j})$ with $\Theta^*_{i,j} \leqslant \rho$ for every $i \in [n]$ and $j \in [m]$. For $\beta = 8/(3N)$, the estimator $\widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}} = \widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}}[\mathfrak{P}]$ satisfies*

$$\frac{\mathbb{E}\big[\|\widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \min_{\mathfrak{p}\in\mathfrak{P}} \Big\{ \inf_{\boldsymbol{\Theta}\in\mathcal{T}[\mathfrak{p}]} \frac{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + 9\sqrt{\frac{\rho}{N}}\, r_{n,m}(K,L) \Big\} + \Big\{ \frac{8\log|\mathfrak{P}|}{3Nnm} \Big\}^{1/2}$$

   provided that $\max_{\mathfrak{p}\in\mathfrak{P}} \big(\frac{1}{m_0}\log(en/n_0) + \frac{1}{n_0}\log(em/m_0)\big) \leqslant 3N\rho$.

2. *(Gaussian model) Assume that $H_{i,j} \sim \mathcal{N}(\Theta^*_{i,j}, \sigma^2_{i,j})$ with $\sigma^2_{i,j} \leqslant \sigma^2$ for every $i \in [n]$ and $j \in [m]$. For $\beta = 4\sigma^2$, the estimator $\widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}} = \widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}}[\mathfrak{P}]$ satisfies*

$$\frac{\mathbb{E}\big[\|\widehat{\boldsymbol{\Theta}}^{\mathsf{EWA}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \min_{\mathfrak{p}\in\mathfrak{P}} \Big\{ \inf_{\boldsymbol{\Theta}\in\mathcal{T}[\mathfrak{p}]} \frac{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + 5\sigma r_{n,m}(K,L) \Big\} + \Big\{ \frac{\log|\mathfrak{P}|}{nm} \Big\}^{1/2}.$$

To prove the first claim of this theorem, it suffices to combine [Dal22, Cor. 4] with Theorem 6. Similarly, the second claim follows from [Dal22, Cor. 2] and Theorem 6. Similar results can be obtained for arbitrary distribution with bounded support and for the Laplace distribution, using Corollary 3 and Corollary 5 from [Dal22], respectively. Unfortunately, we are not aware of

any result that makes it possible to carry these bounds over to the Poisson and the general sub-Gaussian distributions.

Remark 2. The upper bounds obtained in Theorem 7 show that the extra error term due to aggregation is not large, when the sample size nm is large. Note that $\log|\mathfrak{P}|$ is usually not large. A reasonable choice for this set is the following: choose geometrically increasing sequences $K_i = \lfloor 2^{1+i/2} \rfloor$ and $L_j = \lfloor 2^{1+j/2} \rfloor$ for $0 \leqslant i \leqslant 2\log_2(n/10)$ and $0 \leqslant j \leqslant 2\log_2(m/10)$. Then, for each K_i and L_j, choose $n_0 \leqslant n/K_i$ and $m_0 \leqslant m/L_j$ to be of the form $\lfloor 2^{2+\ell/2} \rfloor$. This method of choosing $\mathfrak{P}$ ensures that $|\mathfrak{P}| \leqslant 4\log_2^2(n/7)\log_2^2(m/7)$. Therefore, the term $\log|\mathfrak{P}|$ is, in almost all settings, of smaller order than KL; indeed, one can check that $\log|\mathfrak{P}| \geqslant 12$ implies that $\min(n,m) \geqslant 12 \times 10^4$.

2.3.3 Risk bounds for the graphon estimators $\widehat{W}^{\mathsf{LS}}$ and $\widehat{W}^{\mathsf{EWA}}$

A suitable combination of the risk bounds established in Theorem 6 on the error of estimators of the mean matrix $\boldsymbol{\Theta}^* = \mathbf{E}[\mathbf{H}]$ and of inequality (2.7), allows us to get risk bounds for the graphon estimators $\widehat{W}^{\mathsf{LS}}$ and $\widehat{W}^{\mathsf{EWA}}$. We will focus on two classes, piecewise constant and Hölder continuous graphons, for which the evaluation of the approximation error is provided in Proposition 2. Obviously, using this strategy makes the term $\inf_{\boldsymbol{\Theta}\in\mathcal{T}} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}$—the oracle error—appear in the error bound, where $\mathcal{T}$ is the set of constant by-block-matrices defined in (2.3). In the case of the class of piecewise constant graphons, this oracle error vanishes, whereas in the case of Hölder continuous graphons, it needs to be evaluated, which is done in the next proposition. Note that, unlike in Subsection 2.3.1 and Subsection 2.3.2, we now return to the original framework of random $\boldsymbol{\Theta}^*$ and all the expectations comprise integration with respect to the latent variables $\boldsymbol{U}$ and $\boldsymbol{V}$.

Proposition 3. *Let W^* be α-Hölder continuous, i.e., $W^* \in \mathbb{H}_{\alpha,\mathcal{L}}$ for some $\alpha \in (0,1]$, $\mathcal{L} > 0$. Let $\Theta^*_{i,j} = W^*(U_i, V_j)$ for $i \in [n]$ and $j \in [m]$. Let $n_0 \geqslant 2$, $m_0 \geqslant 2$, $K \leqslant n/n_0$ and $L \leqslant m/m_0$ be four integers. Then, the $n \times m$ matrix $\boldsymbol{\Theta}^*$ with entries $\Theta_{i,j}$ satisfies*

$$\mathbb{E}\Big[\inf_{\boldsymbol{\Theta}\in\mathcal{T}} \frac{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}}\Big]^{1/2} \leqslant \frac{3\mathcal{L}}{2}\Big(\frac{1}{K^\alpha} + \frac{1}{L^\alpha}\Big),$$

where $\mathcal{T} = \mathcal{T}^{K,L}_{n_0,m_0}$ is the set of constant-by-block matrices defined in (2.3).

We have now all the necessary ingredients to state the main results of this chapter, quantifying the error of estimating the graphon. We do this first for the least squares estimator, considering in particular that the parameters K and L of the set $\mathcal{T}$ (on which the minimum in Equation (2.2) is computed) is fixed. We then state the result for the exponentially weighted aggregate.

Theorem 8. *Let $\mathbf{H}$ be a $n \times m$ random matrix satisfying Assumption 3 with some graphon*

$W^* : [0,1]^2 \to [0,\rho]$. *Assume that for some constant* $\sigma > 0$*, conditionally to* $\boldsymbol{U}, \boldsymbol{V}$*, the random variables* $(H_{ij} - \mathbb{E}[H_{ij}|\boldsymbol{U}, \boldsymbol{V}])$ *satisfy the* (σ^2, b)*-Bernstein condition.*

1. *Assume that the graphon* W^* *is* (K, L)*-piecewise constant, meaning that for some integers* $K, L \geqslant 2$ *and for* $0 = a_0 < \ldots < a_K = 1$, $0 = b_1 < \ldots < b_L = 1$*, such that*

$$\Delta^{(K)} := \min_{k\in[K]} |a_k - a_{k-1}| \geqslant \frac{8\log(nK)}{n}, \quad \Delta^{(L)} := \min_{\ell\in[L]} |b_\ell - b_{\ell-1}| \geqslant \frac{8\log(mL)}{m}, \tag{2.10}$$

the function W^* *is constant on each rectangle* $[a_{k-1}, a_k[\times[b_{\ell-1}, b_\ell[$*. Then, the estimator* $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}}$ *with* $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} = \widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K, L]$ *defined by* (2.4) *satisfies*

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant \big(27\sigma^2 + 4b\rho\big)^{1/2}\Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} + \rho\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2},$$

provided that $\psi_{n,m}(\Delta^{(K,L)}) = \frac{6\log(2e/\Delta^{(K)})}{m\Delta^{(L)}} + \frac{6\log(2e/\Delta^{(L)})}{n\Delta^{(K)}} \leqslant (\sigma/b)^2$*.*

2. *Assume that the graphon* W^* *is* α*-Hölder continuous, meaning that* $W^* \in \mathbb{H}_{\alpha,\mathcal{L}}$ *for some* $\alpha \in (0, 1]$ *and* $\mathcal{L} > 0$*. Assume that*[4] *the number of nodes* n, m *satisfy* $n \geqslant m$ *and*

$$\frac{m^{2\alpha+1}}{n\log^4(2n)} \geqslant \mathcal{L}^2\frac{(4b/\sigma)^{4(\alpha+1)} \vee 3}{(25\sigma^2 + 4b\rho)}. \tag{2.11}$$

Let $\beta = \alpha/(2\alpha + 2)$*. Then, there is a choice of* K, L, n_0, m_0 *such that the least squares estimator* $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}}$ *with* $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} = \widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K, L]$ *satisfies*

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant 6\mathcal{L}^{1-2\beta}\Big(\frac{25\sigma^2 + 4b\rho}{3nm}\Big)^{\beta} + \Big(\frac{(50\sigma^2 + 8b\rho)\log m}{m}\Big)^{1/2} + \frac{4\mathcal{L}}{m^{\alpha/2}}. \tag{2.12}$$

In order to ease understanding of these results, let us make some comments. First, one can note that applying the first claim of the theorem to Bernoulli random variables[5] $H_{i,j}$ (the Bernstein condition is then fulfilled with $\sigma^2 = \rho$, $b = 1/3$), we obtain

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant 10\Big(\frac{\rho KL}{nm} + \frac{\rho\log K}{3m} + \frac{\rho\log L}{3n}\Big)^{1/2} + \rho\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2}, \tag{2.13}$$

provided that $\frac{2}{3m\Delta^{(L)}}\log(2e/\Delta^{(K)}) + \frac{2}{3n\Delta^{(K)}}\log(2e/\Delta^{(L)}) \leqslant \rho$. In the balanced setting $n = m$, $K = L$ and $\Delta^{(K)} = \Delta^{(L)} = 1/K$, the upper bound in (2.13) simplifies to

$$10\sqrt{\rho}\,\frac{K}{n} + 9\Big(\frac{\rho\log K}{n}\Big)^{1/2} + \sqrt{2}\rho\Big(\frac{K}{n}\Big)^{1/4},$$

provided that $\rho \geqslant 2(K/n)\log(2eK)$. The last expression is of the same order as the rate

[4] The assumption $n \geqslant m$ does not cause any loss of generality, since n and m play symmetric roles in the framework under consideration.

[5] In fact, exactly the same result holds if the Bernoulli distribution is replaced by any distribution supported by $[0, 1]$.

Distr. of H_{ij}	Values (σ^2, b)	Condition (2.11)	Risk Bound (2.12)
Bernoulli(ρ)	$(\rho, 1/3)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 n \log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(nm)^{1/4}} + \dfrac{8\sqrt{\rho \log m}}{\sqrt{m}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
Binomial$(N, \rho)/N$	$(\rho/N, 1/3N)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 N n \log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Nnm)^{1/4}} + \dfrac{8\sqrt{\rho \log m}}{\sqrt{Nm}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
sub-Gauss(σ^2)	$(\sigma^2, 0)$	$\sigma^2 \geqslant \dfrac{3\mathcal{L}^2 n \log^4(2n)}{25 m^3}$	$\dfrac{11\sqrt{\mathcal{L}\sigma}}{(nm)^{1/4}} + \dfrac{8\sigma\sqrt{\log m}}{\sqrt{m}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
Poisson$(T\rho)/T$	$(\rho/T, 1/3T)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 T n \log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Tnm)^{1/4}} + \dfrac{8\sqrt{\rho \log m}}{\sqrt{Tm}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$

Table 2.1: Upper bound for Lipschitz-continuous graphons and various distributions.

established in [KTV17, Corollary 3.3.i)] for graphons of unipartite graphs. Furthermore, it holds under more general conditions on the observations and contains explicit values for the constants.

Second, one can have a closer look at the order of magnitude of the three terms appearing in (2.12) in the case $n = m$ tending to infinity and assuming $\mathcal{L}, \sigma, b$ and ρ to be of order one. Then the first term is of order $(n^2)^{-\alpha/(2\alpha+2)}$, which is known to be the minimax optimal rate of estimating an α-Hölder continuous, $d = 2$-variate regression function based on n^2 observations. The second term, of order $(\log n/n)^{1/2}$, is dominated by the first term when $\alpha < 1$, and has the optimal order up to a logarithmic factor when $\alpha = 1$. The third term being of order $n^{-\alpha/2}$, is of optimal order $n^{-1/2}$ when $\alpha = 1$, and is the largest term of the sum for all $\alpha < 1$.

To the best of our knowledge, the question of whether there are estimators of Hölder-continuous graphons that achieve a faster rate of convergence than $n^{-\alpha/2}$ remains open. The common belief is that this term is unavoidable and it is the price to pay for not observing the covariates $\boldsymbol{U}, \boldsymbol{V}$. Note that the deterioration caused by this lack of information, measured by the ratio of the third and the first terms of the risk bound in (2.12) is of order $n^{\alpha(1-\alpha)/(2\alpha+2)} \leqslant n^{0.086}$, when n and m are of the same order. From a practical point of view, this deterioration is not significant, since even for $n = 10^9$, $n^{0.086} \leqslant 6$.

One can also draw the consequences of the second claim of the theorem under various (conditional to $\boldsymbol{U}$, $\boldsymbol{V}$) distributions of $H_{i,j}$. For $\alpha = 1$ ($\beta = 1/4$), that is Lipschitz-continuous graphons, conditions (2.11) and inequality (2.12) are reported in Table 2.1.

To close this section, we state the risk bounds that can be obtained by combining Theorem 7 and Theorem 8. To keep the statement simple, only the case of piecewise constant graphon is presented.

Corollary 2. Let $n \leqslant m$ and let $\mathbf{H}$ be a $n \times m$ random matrix satisfying Assumption 3 with some (K, L)-piecewise constant graphon $W^* : [0,1]^2 \to [0, \rho]$. This means that for some integers $K, L \geqslant 2$ and for $0 = a_0 < \ldots < a_K = 1$, $0 = b_1 < \ldots < b_L = 1$ such that (2.10) holds, the function W^* is constant on each rectangle $[a_{k-1}, a_k[\times[b_{\ell-1}, b_\ell[$. Let $\mathfrak{P}$ be chosen as

in Remark 2 and let $\widehat{W}^{\mathsf{EWA}}[\mathfrak{P}] = W_{\widehat{\mathbf{\Theta}}^{\mathsf{EWA}}[\mathfrak{P}]}$.

1. (Bernoulli/binomial model) Assume that for some $N \in \mathbb{N}$, conditionally to $(\boldsymbol{U}, \boldsymbol{V})$,$NH_{i,j}$ is drawn from the $\mathsf{binomial}(N, \Theta^*_{i,j})$ distribution with $\Theta^*_{i,j} = W^*(U_i, V_j) \leqslant \rho$ for every $i \in [n]$ and $j \in [m]$. For $\beta = 8/(3N)$, we have

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{EWA}}, W^*)^2]^{1/2} \leqslant 9\Big(\frac{2\rho KL + \log\log_2 m}{nm} + \frac{\rho \log K}{m} + \frac{\rho \log L}{n}\Big)^{1/2} + \rho\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2}$$

   provided that $\frac{\log(3e/\Delta^{(K)})}{m\Delta^{(L)}} + \frac{\log(3e/\Delta^{(L)})}{n\Delta^{(K)}} \leqslant N\rho$.

2. (Gaussian model) Assume that, conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the entries $H_{i,j}$ are drawn from the Gaussian $\mathcal{N}(W^*(U_i, V_j), \sigma^2(U_i, V_j))$ distribution with $|W^*(u, v)| \leqslant \rho$ and $\sigma^2(u, v) \leqslant \sigma^2$ for every $u, v \in [0, 1]$. For $\beta = 4\sigma^2$, we have

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{EWA}}, W^*)^2]^{1/2} \leqslant 5\sigma\Big(\frac{2KL + \log\log_2 m}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} + 2\rho\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2}.$$

2.4 Tractable approximation of the least-squares estimator

The least squares estimator introduced in (2.2) and studied in previous sections is a solution to a combinatorial optimization problem that is computationally intractable. It is impossible to compute this estimator in polynomial time. The goal of this section is to present a tractable algorithm that computes an approximation of $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$. Of course, there is no guarantee that the presented algorithm provides an estimator that is always close to $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$, but it is plausible that this is true in many cases.

The proposed approximation can be seen as a version of Lloyd's algorithm for k-means clustering [Llo82]. To describe it, let us recall that the least square estimator is defined by

$$(\widehat{\mathbf{Q}}, \widehat{\mathbf{Z}}^{\mathrm{user}}, \widehat{\mathbf{Z}}^{\mathrm{item}})^{\mathsf{LS}} \in \arg \min_{\substack{\mathbf{Q} \in \mathbb{R}^{K \times L} \\ \mathbf{Z}^{\mathrm{user}} \in \mathcal{Z}(n,K,n_0) \\ \mathbf{Z}^{\mathrm{item}} \in \mathcal{Z}(m,L,m_0)}} \big\|\mathbf{H} - \mathbf{Z}^{\mathrm{user}} \mathbf{Q} (\mathbf{Z}^{\mathrm{item}})^\top\big\|_{\mathsf{F}}^2. \tag{2.14}$$

It turns out that when we fix two of the three arguments $\mathbf{Q}, \mathbf{Z}^{\mathrm{user}}, \mathbf{Z}^{\mathrm{item}}$ of $\mathcal{L}$, the minimization problem with respect to the third becomes tractable. We can therefore use the alternating minimization algorithm below, with the guarantee that the cost function $\mathcal{L}(\mathbf{Z}^{\mathrm{user}}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}}) = \|\mathbf{H} - \mathbf{Z}^{\mathrm{user}} \mathbf{Q} (\mathbf{Z}^{\mathrm{item}})^\top\|_{\mathsf{F}}^2$ decreases at each iteration. Different versions of this algorithm have been studied in the literature on estimation and detection in the presence of a latent structure [CO10, LR15, LZ16, GV19].

The rest of this section provides more details on each step of this algorithm, as well as on the initialization and on a stopping criterion.

Algorithm 4 Lloyd's algorithm of alternating minimization for approximating the LSE (2.14)

Require: $\mathbf{Z}^{\mathrm{user}}$, $\mathbf{Z}^{\mathrm{item}}$ the left and right cluster matrices with entries in $\{0,1\}$, $\mathbf{H}$ the data matrix.
Ensure: $(Z_1, \mathbf{Q}, Z_2)$ local minimizer of $\mathcal{L}(\cdot,\cdot,\cdot)$.

Repeat:

1. Compute $\mathbf{Q} = (\mathbf{Z}^{\mathrm{user}}_{\mathsf{norm}})^\top \mathbf{H} \mathbf{Z}^{\mathrm{item}}_{\mathsf{norm}}$ where $\mathbf{Z}^{\mathrm{user}}_{\mathsf{norm}}$ is the matrix $\mathbf{Z}^{\mathrm{user}}$ with normalized columns with respect to ℓ^1-norm (the number of 1 in the column), and similarly for $\mathbf{Z}^{\mathrm{item}}_{\mathsf{norm}}$.
2. Update $\mathbf{Z}^{\mathrm{user}}$ that minimize $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}})$
3. Update $\mathbf{Z}^{\mathrm{item}}$ that minimize $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z}^{\mathrm{user}}, \mathbf{Q}, \mathbf{Z})$

Minimization in the second argument $\mathbf{Q}$ When the clusters are known, meaning that we know matrices $\mathbf{Z}^{\mathrm{user}} \in \mathcal{Z}(n, K, n_0)$ and $\mathbf{Z}^{\mathrm{item}} \in \mathcal{Z}(m, L, m_0)$, see (2.1), the solution to

$$\widehat{\mathbf{Q}} = \arg \min_{\mathbf{Q} \in \mathbb{R}^{K \times L}} \mathcal{L}(\mathbf{Z}^{\mathrm{user}}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}})$$

is easy to compute: each entry $\widehat{Q}_{kl}$ is the average of the coefficients H_{ij} belonging to the (k,l)-block defined by $\mathbf{Z}^{\mathrm{user}}$ and $\mathbf{Z}^{\mathrm{item}}$ (a coefficient (i,j) is in the block (k,l) if $Z^{\mathrm{user}}_{ik} = 1$ and $Z^{\mathrm{item}}_{jl} = 1$). Formally, this is equivalent to $\mathbf{Q} = (\mathbf{Z}^{\mathrm{user}}_{\mathsf{norm}})^\top \mathbf{H} \mathbf{Z}^{\mathrm{item}}_{\mathsf{norm}}$ where $\mathbf{Z}_{\mathsf{norm}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}$ for $\mathbf{Z} \in \{\mathbf{Z}^{\mathrm{user}}, \mathbf{Z}^{\mathrm{item}}\}$.

Minimization with respect to $\mathbf{Z}^{\mathrm{user}}$ We focus now on the problem of minimizing the cost function $\mathcal{L}(\mathbf{Z}^{\mathrm{user}}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}})$ over $\mathbf{Z}^{\mathrm{user}} \in \mathcal{Z}(n, K, n_0)$. Let us first consider the relatively simple case $n_0 = 0$ when there is no constraint on the cardinality of left clusters. We aim to find $\widehat{\mathbf{Z}}^{\mathrm{user}} \in \mathbb{R}^{n \times K}$ that minimizes $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}})$ under the constraints that $\mathbf{Z} \in \{0,1\}^{n \times K}$ and $\mathbf{Z}\mathbf{1}_K = \mathbf{1}_n$ (*i.e.*, each row of $\mathbf{Z}$ has only one entry equal to 1). Let us define $C^{\mathrm{item}}_\ell = \{j \in [m], Z^{\mathrm{item}}_{i\ell} = 1\}$ to be the ℓ-th right cluster and introduce notation

$$\bar{H}^{\mathrm{item}}_{i\ell} = \frac{1}{|C^{\mathrm{item}}_\ell|} \sum_{j \in C^{\mathrm{item}}_\ell} H_{ij} \quad \text{and} \quad \mathbf{D} = \mathsf{diag}(|C^{\mathrm{item}}_\ell|)_{\ell \in [L]}.$$

Simple algebra yields

$$\mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}}) = \sum_{i=1}^{n} \|H_{i,\bullet} - \boldsymbol{Z}^\top_{i,\bullet} \mathbf{Q} (\mathbf{Z}^{\mathrm{item}})^\top\|_2^2,$$

where $\boldsymbol{Z}_{i,\bullet}$ is ith row of $\mathbf{Z}$. Since $\boldsymbol{Z}_{i,\bullet}$ is allowed to have only one nonzero entry, and it should be equal to one, $\boldsymbol{Z}^\top_{i,\bullet} \mathbf{Q} (\mathbf{Z}^{\mathrm{item}})^\top$ is merely equal to one row of $\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top$. This implies that $\widehat{Z}^{\mathrm{user}}_{ij} = \mathbb{1}(j = k_i)$, where

$$\widehat{k}_i = \arg \min_{k \in [K]} \left\|H_{i,\bullet} - \mathbf{Q}_{k,\bullet} (\mathbf{Z}^{\mathrm{item}})^\top\right\|_2^2 = \arg \min_{k \in [K]} \sum_{\ell=1}^{L} \sum_{j \in C^{\mathrm{item}}_\ell} (H_{ij} - Q_{kl})^2.$$

We can rewrite the above expression of $\widehat{k}_i$ as

$$\widehat{k}_i = \arg\min_{k\in[K]} \|\mathbf{D}^{1/2}(\bar{\boldsymbol{H}}^{\mathrm{item}}_{i,\bullet} - \boldsymbol{Q}_{k,\bullet})^\top\|_2^2.$$

Thus, in order to determine $\mathbf{Z}^{\mathrm{user}}$, it suffices to compute the matrix $\bar{\mathbf{H}}^{\mathrm{item}} \in \mathbb{R}^{n\times L}$ and then find for each row of $\bar{\mathbf{H}}^{\mathrm{item}}$ the closest row of $\mathbf{Q}$. Of course, the same procedure is valid for minimizing $\mathcal{L}$ with respect to $\mathbf{Z}^{\mathrm{item}}$ for known $\mathbf{Z}^{\mathrm{user}}$ and $\mathbf{Q}$.

Let us return to the general case $n_0 \geqslant 0$. In this case, we show that the minimization of $\mathcal{L}$ with respect to $\mathbf{Z}^{\mathrm{user}}$ can be done by solving a linear program. Indeed, let is define

$$\phi(\mathbf{Z}) = -2\,\mathrm{tr}(\mathbf{Z}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top\mathbf{H}^\top) + \sum_{k=1}^{K} \mathbf{1}_n^\top \boldsymbol{Z}_{\bullet,k}\boldsymbol{Q}_{k,\bullet}\mathbf{D}\boldsymbol{Q}_{k,\bullet}$$

$$\widetilde{\mathcal{Z}}(n,K,n_0) = \left\{\mathbf{Z}\in[0,1]^{n\times K} : \mathbf{Z}\mathbf{1}_K = \mathbf{1}_n \text{ and } \min_{k\in[K]} \mathbf{1}_n^\top \boldsymbol{Z}_{\bullet,k} \geqslant n_0\right\}.$$

Note that ϕ is a linear function of $\mathbf{Z}^{\mathrm{user}}$, whereas $\widetilde{\mathcal{Z}}(n,K,n_0)$ is a convex polytope containing $\mathcal{Z}(n,K,n_0)$.

Proposition 4. *The following two claims hold true.*

1. *The function $\mathcal{L}(\mathbf{Z},\mathbf{Q},\mathbf{Z}^{item}) - \phi(\mathbf{Z}) = \|\mathbf{H}\|_{\mathsf{F}}^2$ is independent of $\mathbf{Z}$.*
2. *The set of extreme points of $\widetilde{\mathcal{Z}}(n,K,n_0)$ is $\mathcal{Z}(n,K,n_0)$. Equivalently, an element of $\widetilde{\mathcal{Z}}(n,K,n_0)$ is an extreme point if and only if all its entries are either 0 or 1.*

The set $\widetilde{\mathcal{Z}}(n,K,n_0)$ is a convex polytope because it is defined by linear constraints. A well known result [BT97, p 65, Thm 2.7] implies that if $\mathcal{L}(\cdot,\mathbf{Q},\mathbf{Z}^{\mathrm{item}})$ has a minimizer in the polytope $\widetilde{\mathcal{Z}}(n,K,n_0)$, then it has at least one solution in the set of its extreme points $\mathcal{Z}(n,K,n_0)$. There are many efficient solvers for finding such a solution.

Initialization The initialization of Algorithm 4 might have a strong impact on the final result. One possible strategy is to run in parallel N instances of the algorithm with different initialization values, chosen at random. The final estimator is the one that minimizes $\mathcal{L}$ among the resulting (N) candidates.

Another strategy, often used in conjunction with Lloyd's algorithm, is based on spectral initialization. When the graphon is piecewise constant, the problem under consideration is nothing else but the bi-stochastic block model for bipartite networks. Therefore, initial values $(\mathbf{Z}^{\mathrm{user}},\mathbf{Z}^{\mathrm{item}})$ can be obtained, for instance, by the spectral method from [ZA19a]. It consists in computing the K-truncated singular value decomposition of a regularized version of matrix $\mathbf{H}$, and then, in applying k-means clustering to the K-truncated left singular vectors to obtain an initialization for $\mathbf{Z}^{\mathrm{user}}$. The procedure is similar for $\mathbf{Z}^{\mathrm{item}}$.

Stopping rule As mentioned, the cost function is non increasing over the iterations, and it takes its values in a finite set since there is only a finite number of configurations for $(\mathbf{Z}^{\text{user}}, \mathbf{Z}^{\text{item}})$. This is why, from a certain iteration onwards, the values of the cost function remain constant. However, the algorithm may require a large number of iterations to achieve consistency. This suggests to stop iterating if either two consecutive values of the cost function are equal or the maximum number of iterations is attained.

To conclude this section, we stress once again that there is no guarantee that the computationally tractable algorithms we presented here provide the global minimum of the cost function $\mathcal{L}$. However, as can be seen from the numerical examples in Section 2.6, results are quite satisfactory.

2.5 Lower bounds on the minimax risk

We show in this section, that the least squares estimator $\widehat{W}^{\mathsf{LS}}$ is optimal, among all possible estimators, in the sense of its rate of convergence in the worst case over the class

$$\mathcal{W}_\rho[K,L] = \Big\{ W : [0,1]^2 \to [0,\rho] : \exists\, \{I_k\}_{k=1}^K, \{J_\ell\}_{\ell=1}^L \text{ s.t. } W = \sum_{k=0}^{K-1}\sum_{\ell=0}^{L-1} W(a_k,b_\ell)\mathbb{1}_{I_k\times J_\ell} \Big\},$$

where $I_k = [a_k, a_{k+1})$ and $J_\ell = [b_\ell, b_{\ell+1})$ form a partition of $[0,1)$ into intervals. The lower bound will be proven for the binomial model, but all the techniques used in the proof can be extended to the other models presented in the introduction.

Theorem 9. *Assume that conditionally to $(\boldsymbol{U},\boldsymbol{V})$, the entries $H_{i,j}$ of the observed $n\times m$ matrix $\mathbf{H}$ are independent and drawn from the Binomial distribution with parameter $(N, W^*(U_i,V_j))$. There exist universal constants $c, C>0$, such that for any $K,L>C$ satisfying $KL \geqslant L\log^2 L + K\log^2 K$ and for any $\rho>0$,*

$$\inf_{\widehat{W}}\sup_{W^*} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W},W^*)\big]^{1/2} \geqslant c\Big[\sqrt{\rho}\Big(\frac{KL}{Nnm}\wedge\rho + \frac{1}{N\sqrt{nm}}\wedge\rho\Big)^{1/2} + \rho\Big(\sqrt{\frac{K}{n}}+\sqrt{\frac{L}{m}}\Big)^{1/2}\Big], \tag{2.15}$$

where the inf is over all possible estimators $\widehat{W}$ and the sup is over all $W^*\in\mathcal{W}_\rho[K,L]$.

In this theorem, we could assume $NH_{i,j}\sim\mathcal{B}(N,W^*(U_i,V_j))$ – instead of assuming $H_{i,j}\sim\mathcal{B}(N,W^*(U_i,V_j))$) – and obtain the same result. In that sens, the right-hand side of (2.15) should be compared to (2.13). One can observe that if the values of n,m, K, L and ρ are such that the dominating term in the upper bound is one of the terms $\rho(K/n)^{1/4}$ and $\rho(L/m)^{1/4}$, then the lower bound in (2.15) is of the same order as the upper bound. Therefore, in this case, the least squares estimator of the graphon is minimax-rate-optimal. Similarly, if the dominating term is $(\rho KL/(Nnm))^{1/2}$ and $\rho \geqslant KL/(Nnm)$, then the LSE is minimax-rate-optimal. Note also

that if ρ is very small, that is smaller than both $KL/(Nnm)$ and $(N^2nm)^{-1/2}$, then the lower bound in (2.15) is of order ρ, which might be much smaller than the upper bound established for the LSE. This is not an artifact of the proof, but reflects the fact that in this situation the naive estimator $\widehat{W} \equiv 0$ is better than the LSE. Furthermore, this naive estimator turns out to be minimax-rate-optimal under the mentioned constraint on ρ.

Figure 12 depicts the regions of the values of n and m where the lower and the upper bounds are of the same order, illustrating thus the optimality of the least-squares estimator. Similarly, Figure 13 shows the regions of the values of ρ and $\gamma = K/n = L/m$, for fixed values of n and m, where the lower bound is larger than half of the upper bound. We clearly see that even for very unbalanced graphs (m much larger than n), the purple region covers almost the whole square, which means that the least-squares estimator is minimax-rate-optimal in this region.

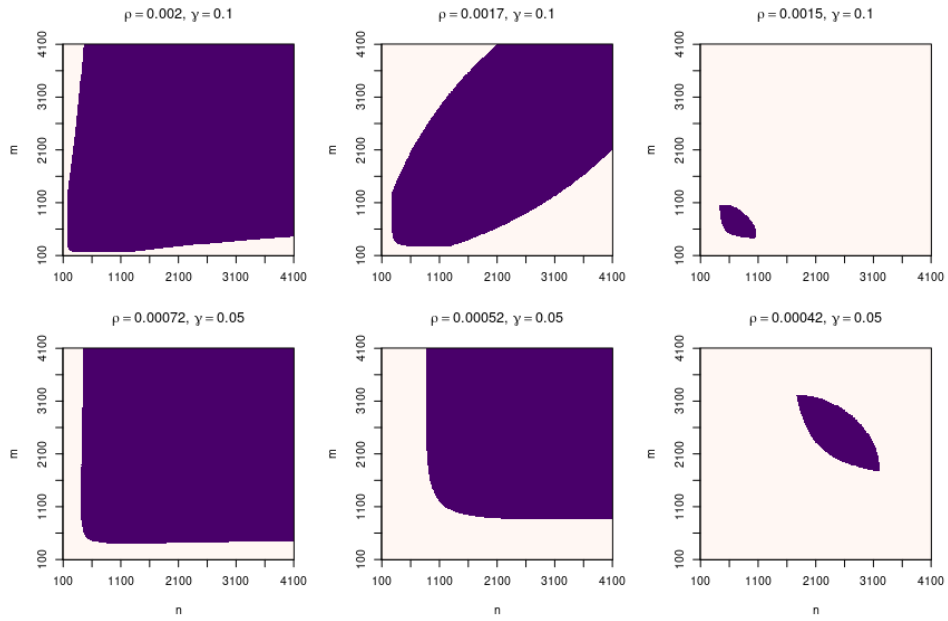


Figure 12: Illustration of the optimality of the least squares estimator for $N = 1$. The purple area corresponds to the values of n and m, for some fixed values of ρ and $\gamma = K/n = L/m$, for which the lower bound is within a constant factor of the upper bound. More precisely, when $\rho\gamma^2 \wedge \rho^2 + \rho(nm)^{-1/2} \wedge \rho^2 + 2\rho^2\sqrt{\gamma}$ is larger than half of $\rho\gamma^2 + (\rho \log K)/(3m) + (\rho \log L)/(3n) + 2\rho^2\sqrt{\gamma}$. We observe that unless ρ is very small, the upper bound established for the least-squares estimator is within a constant factor of the lower bound for all estimators for most values of n and m.

2.6 Numerical experiments

In this section, we present the results of some numerical experiments illustrating the behavior of the error of the estimated graphon $\widehat{W}^{\mathsf{LS}}$ and its dependence on different parameters of the model. We first consider the case of piecewise constant graphons and study the estimation

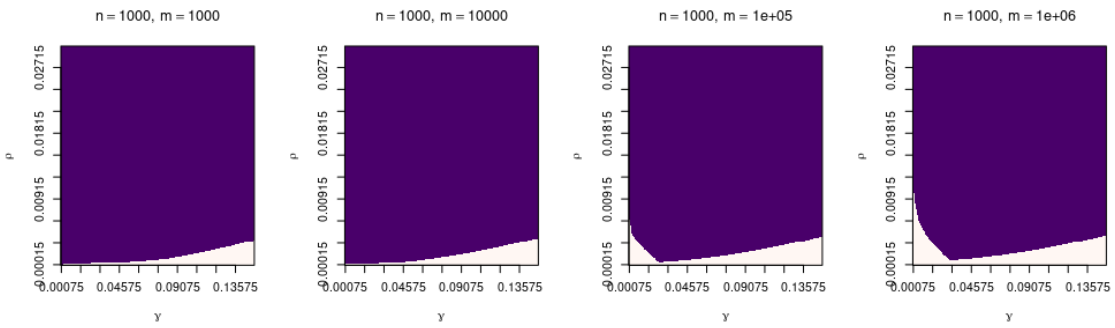


Figure 13: Illutration of the optimality of the least squares estimator. The purple area corresponds to the values of ρ and $\gamma = K/n = L/m$, for some fixed values of n and m, for which the lower bound is within a constant factor of the upper bound. We observe that unless ρ is very small, the upper bound and the lower bound for are of the same order.

error of the matrix $\mathbf{\Theta}^*$. We explore the dependence of this error on n for different values of (ρ, K, L) (assuming that $m = n/2$) as well as on the sparsity parameter ρ for different values of (n, m, K, L). We then show the results of the estimation for a Hölder-continuous graphon, when parameters K and L are chosen as functions of n and m respectively, as recommended by our theoretical results.

2.6.1 Estimation error of the piecewise constant matrix $\mathbf{\Theta}^*$

We report the results of two different experimental set-ups, referred to as rand-graphon and cos-graphon. The two set-ups differ in the choice of the graphon only. In both cases, the partitions on which the graphon is piecewise constant is the regular partition induced by the rectangles of the form $[(k-1)/K, k/K) \times [(\ell-1)/L, \ell/L)$ for $k \in [K]$ and $\ell \in [L]$. In the rand-graphon set-up, the values of W^* are chosen at random between 0 and ρ, while in the cos-graphon set-up, W^* is defined as

$$W^*(u,v) = \frac{2\rho}{3} + \frac{\rho}{3}\cos\big(\pi\lfloor Ku\rfloor\lfloor Lv\rfloor\big), \qquad \forall u, v \in [0,1].$$

Note that the problem is harder than estimating this as a function on $[0,1]^2$ – the usual function estimation setting – because the variables U_i and V_j are not observed. The results obtained in these two set-ups are depicted in Figure 14 and Figures 15 and 16, respectively. In each experiment, we chose $m = n/2$ and computed the median of the squared error $\frac{1}{nm}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2$ for 50 independent repetitions. The estimator $\widehat{\mathbf{\Theta}}$ was computed by Algorithm 4.

For better legibility, the errors in the plots are presented using a log-scale. To check the consistency of the numerical results with our theoretical results, we plotted (in green) the remainder term appearing in the upper-bound in theorem 6. We also displayed the oracle error (red curve) corresponding to the error of the best pseudo-estimator that is built with the knowledge of the true left and right cluster matrices. We only computed the block averages

in this case. The labels “spectral” and “random” refer to the initialization process used for the algorithm. To display the uncertainty, we plotted colored areas corresponding to the quantiles of order 0.1 and 0.9 respectively. (One may be surprised by the fact that this area grows with n in some cases; this is an artifact of the log-scale). In Algorithm 4, we chose $\gamma = 10^{-3}$ and the maximum number of iterations equal to 40.

One can observe in these experimental results that the error of the “spectral” version always decreases with n and gets closer to the oracle error faster when ρ is larger, as well as when n/K and m/L are higher, following in that the intuition. Indeed, the bigger ρ is, the more links there are, rendering the estimation more accurate. Similarly, the higher n/K and m/L are, making again the estimation more accurate. The “random” version of the algorithm has a more chaotic behavior. It is in most cases larger than the error of the “spectral” version when n/K and m/L are large enough.

In the set-up of cos-graphon, we displayed in Figure 14 the behavior of the error as a function of ρ. For small values of ρ, the random initialization appears to be better than the spectral one. Moreover, the error is increasing for $\rho \in [0, 1/2]$ for both initializations. The reason for such a behavior is that the estimator we computed tries to mimic the oracle estimator, which knows the clusters and estimates the matrix $\mathbf{\Theta}^*$ by computing cluster-wise averages. Thus, if $\mathbf{\Theta}^* = (\mathbf{Z}^{\text{user}})^* \mathbf{Q}^* (\mathbf{Z}^{\text{item}})^*$, where matrices $(\mathbf{Z}^{\text{user}})^*$ and $(\mathbf{Z}^{\text{item}})^*$ represent the clusters, then the oracle $\widehat{\mathbf{\Theta}}_o = (\mathbf{Z}^{\text{user}})^* \widehat{\mathbf{Q}}_o (\mathbf{Z}^{\text{item}})^*$ satisfies

$$\begin{aligned} \frac{\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}_o - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]}{nm} &= \frac{1}{nm} \sum_{k,l} Q^*_{kl}(1 - Q^*_{kl}) \\ &= \frac{1}{nm} \sum_{k,l} \rho \widetilde{Q}_{kl}(1 - \rho \widetilde{Q}_{kl}) \\ &= \frac{\rho(\|\widetilde{\mathbf{Q}}\|_{1,1} - \rho \|\widetilde{\mathbf{Q}}\|_{\mathsf{F}}^2)}{nm} \end{aligned} \tag{2.16}$$

In the above formula, we used the matrix $\widetilde{\mathbf{Q}} = \mathbf{Q}^*/\rho$, which has all its entries in $[0, 1]$, and denoted by $\|\widetilde{\mathbf{Q}}\|_{1,1}$ the sum of entries of the matrix $\widetilde{\mathbf{Q}}$. The right-hand side of (2.16) is a function of ρ that increases for $\rho \in [0, \|\widetilde{\mathbf{Q}}\|_{1,1}/2\|\widetilde{\mathbf{Q}}\|_{\mathsf{F}}^2]$, and decreases outside this interval. Recall that $\widetilde{\mathbf{Q}}$ has all its entries in the interval $[0, 1]$. As a consequence, $\frac{|\widetilde{\mathbf{Q}}|_{1,1}}{|\widetilde{\mathbf{Q}}|^2 \mathsf{F}} \geqslant 1$, which gives an intuition about the increasing behavior of the error for $\rho \in [0, 1/2]$.

2.6.2 Estimation error for Hölder-continuous graphons

To illustrate the behavior of the estimator of the graphon in the case where the latter is Hölder-continuous, we consider the function displayed in Figure 18 and given by

$$W^*(u, v) = \frac{\rho}{2}\Big(1 + \exp\big\{ -10\big((u - 1/2)^2 + (v - 1/2)^2\big)\big\}\Big).$$

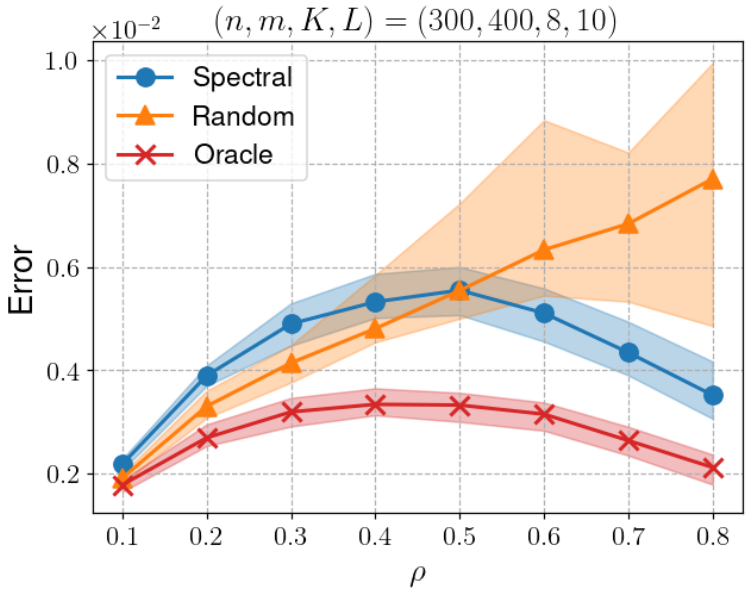


Figure 14: Estimation error as a function of ρ, for $W^*(x,y) = \frac{2\rho}{3} + \frac{\rho}{3}\cos(3\pi\lfloor Kx\rfloor\lfloor Ly\rfloor)$.

This function being Lipschitz-continuous, we have $\alpha = 1$.

The average squared error of estimation over 50 repetitions for different values of n and ρ is depicted in Figure 17. Since the true distance δ is prohibitively hard to compute (because of the minimization over all measure preserving bijections), we computed an approximation of it denoted by $\widetilde{\delta}$. Roughly speaking, $\widetilde{\delta}$ is obtained from δ by replacing the minimum over all bijections τ_1, τ_2 by the value of the cost function at the particular instances of bijections, τ_{σ_1} and τ_{σ_2}, used in the proof of Proposition 2. More precisely, if σ_1 and σ_2 are permutations of $[n]$ and $[m]$, respectively, such that the sequences $(U_{\sigma_1^{-1}(i)})_{i\in[n]}$ and $(V_{\sigma_2^{-1}(j)})_{j\in[m]}$ are nondecreasing, then

$$\tau_{\sigma_1}(u) = \sum_{i=1}^{n}\Big(\frac{\sigma_1(i)-1}{n} - u - \frac{i-1}{n}\Big)\mathbb{1}_{[i-1,i)}(nu), \qquad \tau_{\sigma_1}(1) = \frac{\sigma_1(n)}{n},$$
$$\tau_{\sigma_2}(v) = \sum_{j=1}^{m}\Big(\frac{\sigma_2(j)-1}{m} - v - \frac{j-1}{m}\Big)\mathbb{1}_{[j-1,j)}(nu), \qquad \tau_{\sigma_2}(1) = \frac{\sigma_2(m)}{m}.$$

Then, we define

$$\begin{aligned}\widetilde{\delta}^2(\widehat{W}, W^*) &= \big\|W^* \circ (\tau_{\sigma_1}\otimes\tau_{\sigma_2}) - \widehat{W}\big\|_{\mathbb{L}_2}^2 \\ &= \big\|W^*\big\|_{\mathbb{L}_2}^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{\Theta}_{ij}\int_{\frac{\sigma_1(i)-1}{n}}^{\frac{\sigma_1(i)}{n}}\int_{\frac{\sigma_2(j)-1}{m}}^{\frac{\sigma_2(j)}{m}} W^*(x,y)dxdy + \|\widehat{\boldsymbol{\Theta}}\|_{\mathsf{F}}^2.\end{aligned}$$

In numerical experiments, the integrals appearing in the right-hand side of the last display are approximated by the Riemann sums. In this case also we observe that the error curves

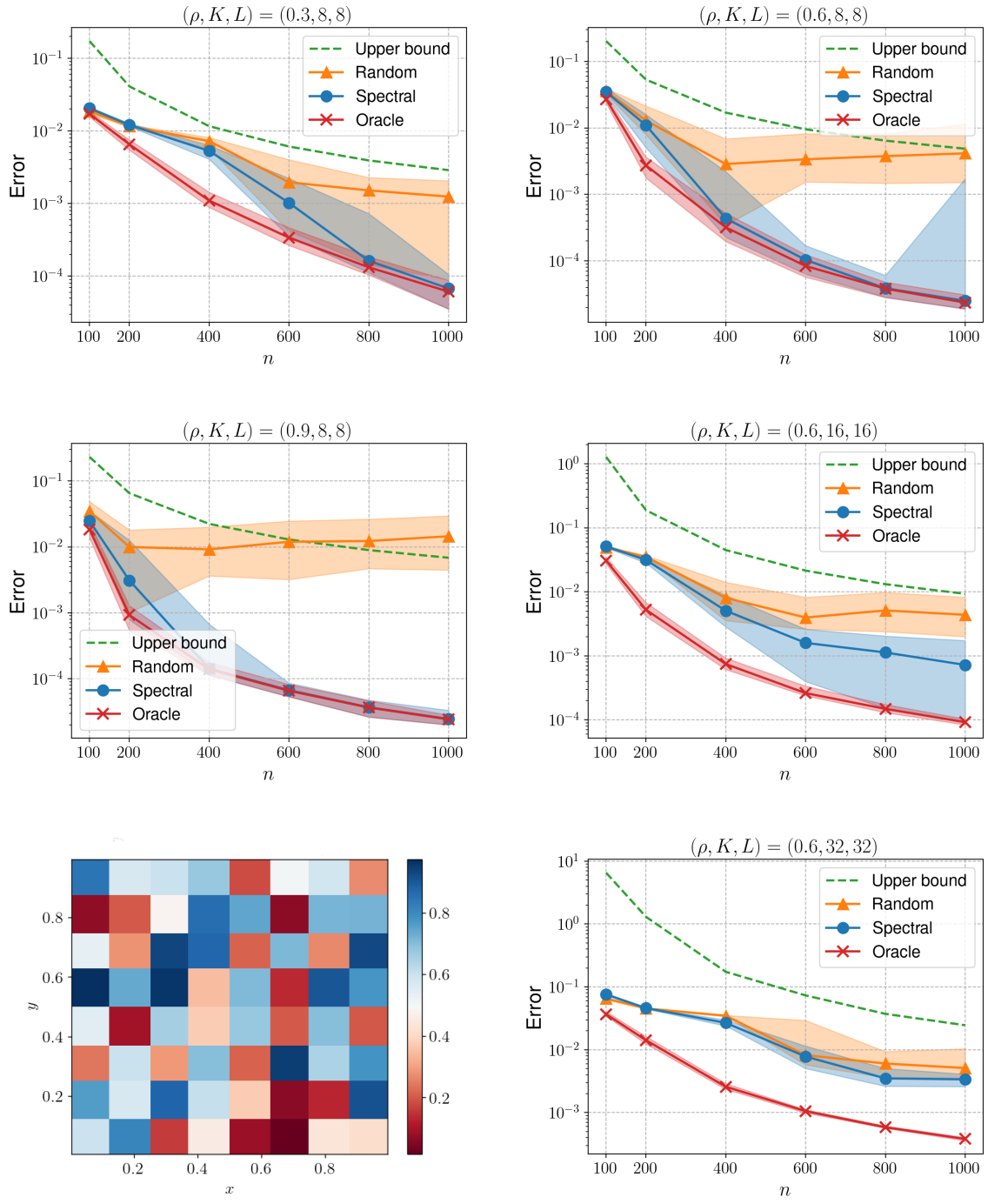


Figure 15: Graphon estimation in the rand-graphon set-up. The random graphon is represented in the figure at bottom left. The others figures plot the error of our pseudo-estimator for different settings.

obtained by Monte Carlo simulations are of the same shape as those predicted by the theory. Interestingly, and somewhat surprisingly, the random initialization behaves as well as the spectral one. We do not have any explanation for this observation at this stage.

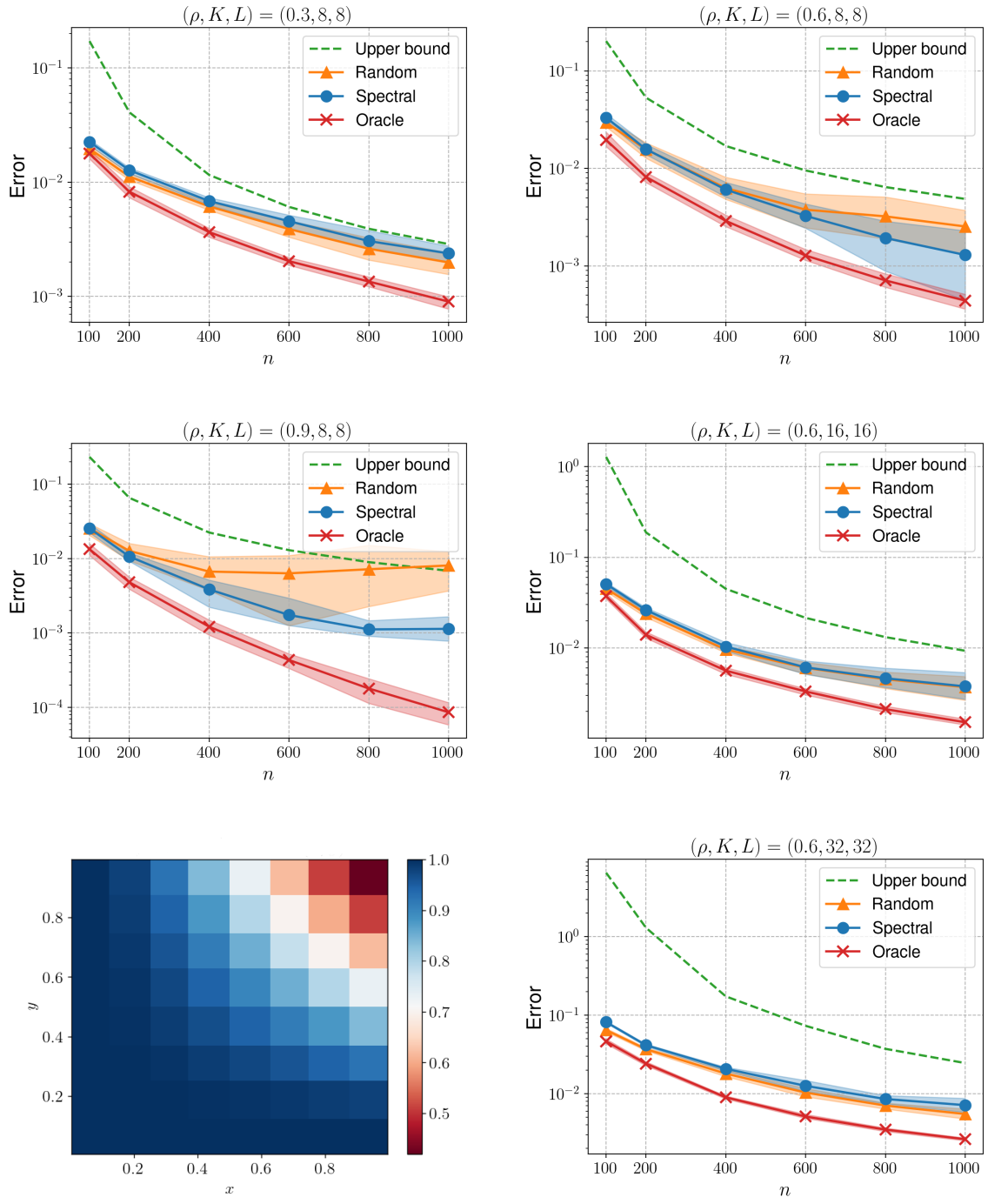


Figure 16: Graphon estimation in the cos-graphon set-up, $W^*(x, y) = \frac{2\rho}{3} + \frac{\rho}{3}\cos\left(3\pi\lfloor Kx\rfloor\lfloor Ly\rfloor\right)$. The graphon is represented in the figure at bottom left. The others figures plot the error of our pseudo-estimator for different settings.

2.7 Proofs of results stated in previous sections

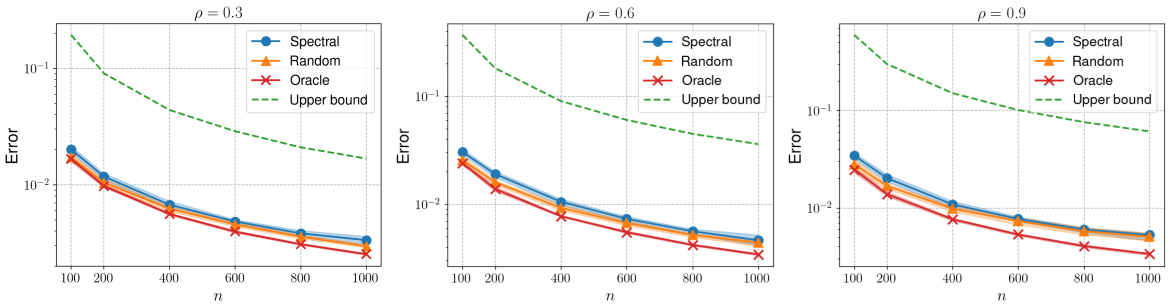
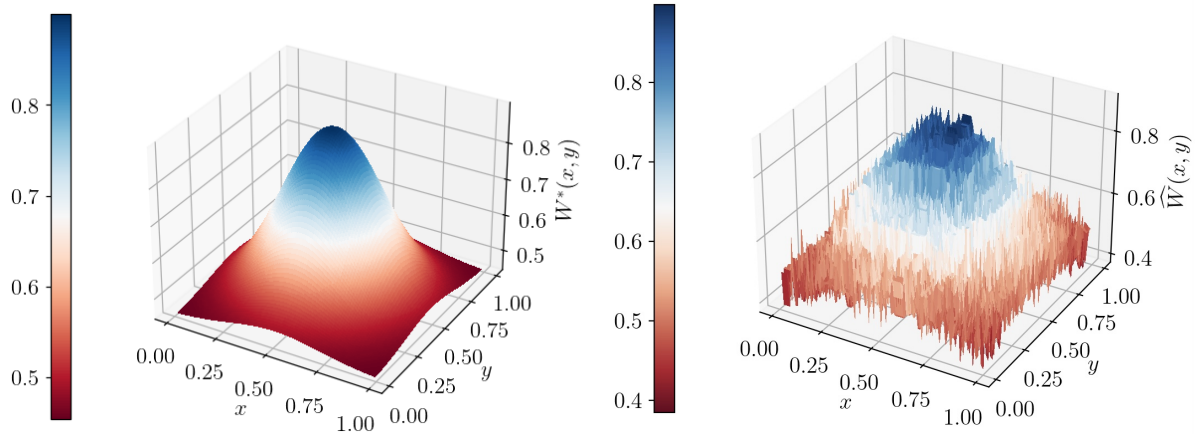


Figure 17: Evolution of the estimation error for a Lipschitz graphon, $m = n/2$.



(a) True graphon W^*

(b) Estimated graphon $\widehat{W}$

Figure 18: Graphon representation for $\rho = 0.9$.

2.7.1 Proof of Theorem 6 (risk bound for $\widehat{\Theta}^{\mathrm{LS}}$)

Let us define

$$\Pi_{\mathcal{T}}(\mathbf{\Theta}^*) = \arg\min_{\mathbf{\Theta} \in \mathcal{T}} \|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}},$$

the best approximation of $\mathbf{\Theta}^*$ in Frobenius norm by a constant-by-block matrix. Note that the matrix $\Pi_{\mathcal{T}}(\mathbf{\Theta}^*)$ has at most KL distinct entries each of which is the average of the entries of a submatrix of $\mathbf{\Theta}^*$. Since $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{Z}}^{\mathrm{user}} \widehat{\mathbf{Q}} (\widehat{\mathbf{Z}}^{\mathrm{item}})^\top$ is the least square estimator, we have

$$\left\|\mathbf{H} - \widehat{\mathbf{\Theta}}\right\|_{\mathsf{F}}^2 \leqslant \left\|\mathbf{H} - \Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\right\|_{\mathsf{F}}^2. \tag{2.17}$$

Let us define the mean-zero "noise" matrix $\mathbf{E} = \mathbf{H} - \mathbb{E}[\mathbf{H}] = \mathbf{H} - \mathbf{\Theta}^*$ and rewrite (2.17) in the following form

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \leqslant \|\mathbf{\Theta}^* - \Pi_{\mathcal{T}}(\mathbf{\Theta}^*)'.\|_{\mathsf{F}}^2 + 2\langle \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*, \mathbf{E}\rangle + 2\langle \mathbf{\Theta}^* - \Pi_{\mathcal{T}}(\mathbf{\Theta}^*), \mathbf{E}\rangle. \tag{2.18}$$

The expectation of $\mathbf{E}$ being zero, the same is true for the last term in the right-hand. We want to bound the expectation of $\langle \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*, \mathbf{E}\rangle$. To this end, we define

$$\widehat{\mathcal{T}} = \Big\{\mathbf{\Theta} : \exists \mathbf{Q} \in \mathbb{R}^{K\times L} \text{ such that } \mathbf{\Theta} = \widehat{\mathbf{Z}}^{\text{user}}\mathbf{Q}(\widehat{\mathbf{Z}}^{\text{item}})^\top\Big\} \subset \mathcal{T},$$

and let $\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) = \arg\min_{\mathbf{\Theta}\in\widehat{\mathcal{T}}} \|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}}$ be the best Frobenius approximation of $\mathbf{\Theta}^*$ in $\widehat{\mathcal{T}}$. We use the decomposition

$$\langle \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*, \mathbf{E}\rangle = \underbrace{\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*, \mathbf{E}\rangle}_{\Xi_1} + \underbrace{\langle \widehat{\mathbf{\Theta}} - \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*), \mathbf{E}\rangle}_{\Xi_2}. \tag{2.19}$$

Lemma 1. *Under the conditions of Theorem 6, we have*

$$\mathbb{E}(\Xi_1) \leqslant \sigma\sqrt{2(n\log K + m\log L + 1)}\,\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2} + b\rho(n\log K + m\log L + 1).$$

Proof. The main steps of the proof consist in applying the Bernstein inequality to Ξ_1 for a fixed $\mathbf{\Theta}$ instead of $\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*)$, using the union bound and then integrating the high-probability bound. For the first step, let $\mathbf{\Theta} \in \mathbb{R}^{n\times m}$ satisfy $\Theta_{i,j} \in [0, \rho]$ for every $(i, j) \in [n] \times [m]$. By definition of the inner product, we have $\langle \mathbf{\Theta} - \mathbf{\Theta}^*, \mathbf{E}\rangle = \sum_{(i,j)\in[n]\times[m]} (\mathbf{\Theta} - \mathbf{\Theta}^*)_{ij} E_{ij}$. The random variables E_{ij} are independent and satisfy the (σ^2, b)-Bernstein condition. The nm-vector with entries $(\mathbf{\Theta} - \mathbf{\Theta}^*)_{ij}$ has an infinity norm bounded by ρ. Therefore, the version of the Bernstein inequality stated in Lemma 12 implies that for all $x > 0$, we have

$$\mathbb{P}\Big(\langle \mathbf{\Theta} - \mathbf{\Theta}^*, \mathbf{E}\rangle \geqslant \sqrt{2x}\,\sigma\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}} + b\rho x\Big) \leqslant e^{-x}.$$

Let us define $\Omega_{\mathbf{Z},\mathbf{Z}'} = \{(\widehat{\mathbf{Z}}^{\text{user}}, \widehat{\mathbf{Z}}^{\text{item}}) = (\mathbf{Z}, \mathbf{Z}')\}$, for each pair of matrices $\mathbf{Z} \in \mathcal{Z}_{n,K,n_0}$ and $\mathbf{Z}' \in \mathcal{Z}_{m,L,m_0}$. On the event $\Omega_{\mathbf{Z},\mathbf{Z}'}$, the matrix $\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*)$ is deterministic and its elements are averages of the elements of $\mathbf{\Theta}^*$. Hence, $0 \leqslant (\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*))_{ij} \leqslant \|\mathbf{\Theta}^*\|_\infty$ and

$$\mathbb{P}\left(\left\{\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*, \mathbf{E}\rangle \geqslant \sqrt{2x}\,\sigma\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*\|_{\mathsf{F}} + b\rho x\right\} \cap \Omega_{\mathbf{Z},\mathbf{Z}'}\right) \leq e^{-x}.$$

Note also that the cardinality of $\mathcal{Z}_{n,K,n_0}$ is at most K^n. Combining the last display with the union bound, we get

$$\begin{aligned}\mathbb{P}\Big(\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*, \mathbf{E}\rangle &\geqslant \sqrt{2x\rho}\,\sigma\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*\|_{\mathsf{F}} + b\rho\sigma^2 x\Big)\\ &\leqslant \sum_{(\mathbf{Z},\mathbf{Z}')} \mathbb{P}\left(\left\{\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*, \mathbf{E}\rangle \geqslant \sqrt{2x}\,\sigma\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*\|_{\mathsf{F}} + b\rho x\right\} \cap \Omega_{\mathbf{Z},\mathbf{Z}'}\right)\end{aligned}$$

$$\leqslant K^n L^m e^{-x},$$

where in the first inequality in the above display the sum is over $(\mathbf{Z}, \mathbf{Z}')$ from the set $\mathcal{Z}_{n,K,n_0} \times \mathcal{Z}_{m,L,m_0}$ and the factor $K^n L^m$ corresponds to an upper bound on the cardinality of this set. Finally, choosing $x = n \log K + m \log L + t$ for some $t > 0$ and using the basic inequality $uv \leqslant \lambda u^2 + v^2/(4\lambda)$ entails

$$\mathbb{P}\left(\Xi_1 \geqslant \lambda \|\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 + \Big(\frac{\sigma^2}{2\lambda} + b\rho\Big)(n \log K + m \log L + t)\right) \leqslant e^{-t}$$

for any $\lambda > 0$. Lemma 11 below ensures that

$$\begin{aligned}\mathbb{E}(\Xi_1) &\leqslant \lambda \mathbb{E}\Big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\Big] + \Big(\frac{\sigma^2}{2\lambda} + b\rho\Big)(n \log K + m \log L + 1) \\ &\leqslant \lambda \mathbb{E}\Big[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\Big] + \Big(\frac{\sigma^2}{2\lambda} + b\rho\Big)(n \log K + m \log L + 1).\end{aligned}$$

Optimizing with respect to $\lambda > 0$, we get

$$\mathbb{E}(\Xi_1) \leqslant \sigma\sqrt{2(n \log K + m \log L + 1)}\, \mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2} + b\rho(n \log K + m \log L + 1).$$

This completes the proof of the lemma. □

We now switch to the evaluation of $\mathbb{E}(\Xi_2)$. To this end, we first notice that

$$\mathbb{E}(\Xi_2) = \mathbb{E}\big[\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{H}) - \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*), \mathbf{E}\rangle\big] = \mathbb{E}\big[\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{E}), \mathbf{E}\rangle\big] = \mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big].$$

Lemma 2. *Under the conditions of Theorem 6, we have*

$$\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big] \leqslant 4(b + \sigma^2)\big(3KL + n \log K + m \log L\big)\big(2\rho + b\psi_{n,m}(n_0, m_0)\big).$$

Proof. Recall that $\psi_{n,m}(n_0, m_0) = \frac{3 \log(en/n_0)}{m_0} + \frac{3 \log(em/m_0)}{n_0}$. The scheme of the proof is to apply successively Lemma 14 and Lemma 15. We will proceed by vectorizing the matrices in order to work with vectors only. To this end, let us consider an arbitrary bijection

$$\phi : [n] \times [m] \to [nm]\,.$$

Let $\mathcal{N}$ and $\mathcal{M}$ be partitions of $[n]$ and $[m]$, respectively, satisfying

$$|\mathcal{N}| = K, \quad |\mathcal{M}| = L \qquad \text{and} \qquad \min_{A \in \mathcal{N}} |A| \geqslant n_0 \quad \min_{B \in \mathcal{M}} |B| \geqslant m_0.$$

We define $\mathcal{N} \times \mathcal{M} = \{A \times B : A \in \mathcal{N}, B \in \mathcal{M}\}$ which is a partition of $[n] \times [m]$ of cardinality $|\mathcal{N} \times \mathcal{M}| = KL$ satisfying $|A \times B| \geqslant n_0 m_0$ for every $A \times B \in \mathcal{N} \times \mathcal{M}$. We denote by $\mathcal{G}$ the family of all partitions $\phi(\mathcal{N} \times \mathcal{M})$, where $\mathcal{N}$ and $\mathcal{M}$ are as above.

Since the entries of $\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})$ are averages of coefficients of $\mathbf{E}$, if we vectorize $\mathbf{E}$ according to the map ϕ, meaning that we define $(\vec{\mathbf{E}})_i = E_{\phi^{-1}(i)}$ for all $i \in [nm]$, we have that

$$\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2 \leqslant \max_{G\in\mathcal{G}} \|\Pi_G \vec{\mathbf{E}}\|_2^2$$

with entries $(\vec{\mathbf{E}})_i$ satisfying the assumptions of Lemma 14. So by the union bound on $\mathcal{G}$, we obtain that

$$\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2 \leqslant 4\big(t + \log(2M|\mathcal{G}|)\big)\big(2\sigma^2 + b \max_{G\in\mathcal{G}} \|\Pi_G \vec{\mathbf{E}}\|_\infty\big) \tag{2.20}$$

with probability at least $1 - 0.5e^{-t}$, where $\log M \leqslant KL \log 12 \leqslant 2.5KL$.

Let $\mathcal{A}$ be the family of all the cells of the partitions in $\mathcal{G}$, and let $\mathcal{A}_{(s,l)} = \{\phi(A \times B) \in \mathcal{A} : |A| = s \text{ and } |B| = l\}$ for $j = (s,l) \in [n] \times [m]$. Define

$$\begin{aligned}
\mathfrak{F} &= \frac{\log(4nm)}{n_0 m_0} + \max_{s,l} \frac{\log|\mathcal{A}_{(s,l)}|}{sl} \\
&\geqslant \frac{\log(4nm)}{n_0 m_0} + \frac{\log|\mathcal{A}_{(n_0,m_0)}|}{n_0 m_0} \\
&\geqslant \frac{7}{n_0 m_0}, \qquad \forall n, m \geqslant 3.
\end{aligned}$$

According to Lemma 15, on an event of probability at least $1 - 0.5e^{-t}$,

$$\begin{aligned}
\max_{G\in\mathcal{G}} \|\Pi_G \vec{\mathbf{E}}\|_\infty &\leqslant \max_{A\in\mathcal{A}} \frac{1}{|A|} \Big| \sum_{\ell\in A} (\vec{\mathbf{E}})_\ell \Big| \\
&\leqslant \sigma \sqrt{2\Big(\frac{t}{n_0 m_0} + \mathfrak{F}\Big)} + b\Big(\frac{t}{n_0 m_0} + \mathfrak{F}\Big) \\
&\leqslant \sigma \sqrt{2\mathfrak{F}(1 + t/7)} + b\mathfrak{F}(1 + t/7) \\
&\leqslant \sigma \sqrt{2\mathfrak{F}}\,(1 + t/14) + b\mathfrak{F}(1 + t/7).
\end{aligned} \tag{2.21}$$

Finally, combining (2.20) and (2.21), we have that with probability at least $1 - e^{-t}$,

$$\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2 \leqslant 4\big(t + \log(2M|\mathcal{G}|)\big)\big(2\sigma^2 + b\sigma\sqrt{2\mathfrak{F}}\,(1 + t/14) + b^2\mathfrak{F}(1 + t/7)\big).$$

Using Lemma 11, we obtain the following upper bound on the expectation

$$\begin{aligned}
\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big] &\leqslant 4\big(1 + \log(2M|\mathcal{G}|)\big)\big(2\sigma^2 + b\sigma\sqrt{2\mathfrak{F}}\,(1 + 1/7) + b^2\mathfrak{F}(1 + 2/7)\big) \\
&\leqslant \big(1 + \log(2M|\mathcal{G}|)\big)\big(9\sigma^2 + 16\mathfrak{F}b^2\big).
\end{aligned}$$

For any (s,l) such that $n_0 \leqslant s \leqslant n$ and $m_0 \leqslant l \leqslant m$, we have $|\mathcal{A}_{(s,l)}| \leqslant \binom{n}{s}\binom{m}{l}$, which implies

$$|\mathcal{A}_{(s,l)}| \leqslant \Big(\frac{en}{s}\Big)^s \Big(\frac{em}{l}\Big)^l.$$

Therefore, taking the logarithm of the two sides, we get

$$\begin{aligned}\frac{\log|\mathcal{A}_{(s,l)}|}{sl} &\leqslant \frac{\log(en/s)}{l} + \frac{\log(em/l)}{s} \\ &\leqslant \frac{\log(en/n_0)}{m_0} + \frac{\log(em/m_0)}{n_0}\,.\end{aligned}$$

This implies that[6] $\mathfrak{F} \leqslant (1/2)\psi_{n,m}(n_0, m_0)$. Therefore,

$$\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big] \leqslant \big(1 + \log(2M|\mathcal{G}|)\big)\big(9\sigma^2 + 8b^2\psi_{n,m}(n_0, m_0)\big)$$

where $|\mathcal{G}| = |\mathcal{F}| \leqslant K^n L^m$ and $M \leqslant 12^{KL}$. Taking into account the fact that $K \geqslant 2$ and $L \geqslant 2$, this leads to

$$\begin{aligned}1 + \log(2M|\mathcal{G}|) &\leqslant 1 + \log 2 + KL\log 12 + n\log K + m\log L \\ &\leqslant 3KL + n\log K + m\log L.\end{aligned}$$

The term $\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big]$ is eventually bounded as follows:

$$\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big] \leqslant \big(3KL + n\log K + m\log L\big)\big(9\sigma^2 + 8b^2\psi_{n,m}(n_0, m_0)\big).$$

This completes the proof of the lemma. □

In order to ease notation in the rest of the proof, let us set $A = n\log K + m\log L$. To conclude, we use the bounds on Ξ_1 and Ξ_2 obtained in Lemma 1 and Lemma 2, respectively, as well as decompositions (2.18) and (2.19). Since $\mathbb{E}[\langle \boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}(\boldsymbol{\Theta}^*), \mathbf{E}\rangle] = \langle \boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}(\boldsymbol{\Theta}^*), \mathbb{E}[\mathbf{E}]\rangle = 0$, we arrive at

$$\begin{aligned}\mathbb{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2] &\leqslant \|\boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}\boldsymbol{\Theta}^*\|_{\mathsf{F}}^2 + 2\mathbb{E}[\Xi_1] + 2\mathbb{E}[\Xi_2] \\ &\leqslant \|\boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}\boldsymbol{\Theta}^*\|_{\mathsf{F}}^2 + 2\sigma\sqrt{2(A+1)}\,\mathbb{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2]^{1/2} + 2b\rho(A+1) + 2\mathbb{E}[\Xi_2].\end{aligned}$$

One can check that the last inequality leads to

$$\big(\mathbb{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2]^{1/2} - \sigma\sqrt{2(A+1)}\big)^2 \leqslant \|\boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}(\boldsymbol{\Theta}^*)\|_{\mathsf{F}}^2 + 2(A+1)(\sigma^2 + b\rho) + 2\mathbb{E}[\Xi_2].$$

This readily yields

$$\begin{aligned}\mathbb{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2]^{1/2} &\leqslant \|\boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}(\boldsymbol{\Theta}^*)\|_{\mathsf{F}} + \sigma\sqrt{2(A+1)} + \big(2(A+1)(\sigma^2 + b\rho) + 2\mathbb{E}[\Xi_2]\big)^{1/2} \\ &\leqslant \|\boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}(\boldsymbol{\Theta}^*)\|_{\mathsf{F}} + \big((A+1)(8\sigma^2 + 4b\rho) + 4\mathbb{E}[\Xi_2]\big)^{1/2} \\ &\leqslant \|\boldsymbol{\Theta}^* - \Pi_{\mathcal{T}}(\boldsymbol{\Theta}^*)\|_{\mathsf{F}} + \big(17\sigma^2 + 4b\rho + 8b^2\psi_{n,m}(n_0, m_0)\big)^{1/2}(3KL + A)^{1/2},\end{aligned}$$

[6] This is true since $\frac{\log(2n)}{n_0\log(en/n_0)} = \frac{1}{n_0} + \frac{\log(2n_0/e)}{n_0\log(en/n_0)} \leqslant \frac{\log(4n_0)}{n_0\log(2e)} \leqslant 0.5$ provided that $n_0 \geqslant 3$.

where in the second line we have used the inequality $\sqrt{x}+\sqrt{y} \leqslant \sqrt{2x+2y}$. Finally, under the condition $\psi_{n,m}(n_0, m_0) \leqslant (\sigma/b)^2$, we get the claim of the proposition.

2.7.2 Proof of Proposition 2 (approximation error for a graphon)

First claim (piecewise constant graphon) In what follows, λ refers to the Lebesgue measure on $\mathbb{R}$ and λ_2 is the Lebesgue measure on $\mathbb{R}^2$. Let W^* be a graphon such that for some $K \times L$ matrix $\mathbf{Q}^*$ and some sequences $a_0 < \ldots < a_K$, $b_0 < \ldots < b_L$ satisfying $a_0 = b_0 = 0$ and $a_K = b_L = 1$, we have $W^*(u,v) = Q^*_{k,\ell}$ for every $u \in [a_{k-1}, a_k)$ and $v \in [b_{\ell-1}, b_\ell)$. Equivalently,

$$W^*(u,v) = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \mathbf{Q}^*_{k,\ell} \mathbb{1}_{[a_{k-1},a_k)\times[b_{\ell-1},b_\ell)}(u,v).$$

Let us also define the "weight" sequences $w^{(1)}_k = a_k - a_{k-1}$, $w^{(2)}_\ell = b_\ell - b_{\ell-1}$ and

$$\widehat{w}^{(1)}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[a_{k-1},a_k[}(U_i) \qquad \text{and} \qquad \widehat{w}^{(2)}_\ell = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{[b_{\ell-1},b_\ell[}(V_j).$$

Notice that all the four weight sequences $w^{(1)}$, $w^{(2)}$, $\widehat{w}^{(1)}$ and $\widehat{w}^{(2)}$ are positive and sum to one. As proved in [KTV17, p16], there exist two functions $\psi_1 : [0,1] \to [K]$ and $\psi_2 : [0,1] \to [L]$ such that

1. For all $k \in [K]$ and $x \in \big[a_{k-1}, (a_{k-1} + \widehat{w}^{(1)}_k) \wedge a_k\big)$, we have $\psi_1(x) = k$
2. For all $\ell \in [L]$ and $x \in \big[b_{\ell-1}, (b_{\ell-1} + \widehat{w}^{(2)}_\ell) \wedge b_\ell\big)$, we have $\psi_2(x) = \ell$
3. $\lambda(\psi_1^{-1}(k)) = \widehat{w}^{(1)}_k$ for all $k \in [K]$
4. $\lambda(\psi_2^{-1}(\ell)) = \widehat{w}^{(2)}_\ell$ for all $\ell \in [L]$.

Using these mappings ψ_1 and ψ_2, we construct the graphon $W^*_\psi(u,v) = \mathbf{Q}^*_{\psi_1(u),\psi_2(v)}$ which satisfies $\delta(W_{\mathbf{\Theta}^*}, W^*_\psi) = 0$. This leads to

$$\begin{aligned}\delta^2(W_{\mathbf{\Theta}^*}, W^*) = \delta^2(W^*_\psi, W^*) &\leqslant \|W^*_\psi - W^*\|^2_{\mathbb{L}_2} \\ &\leqslant (B-A)^2 \lambda_2\big((u,v) : W^*_\psi(u,v) \neq W^*(u,v)\big).\end{aligned}$$

Our choice of W^*_ψ ensures that $W^*(u,v) = W^*_\psi(u,v)$ except if $u \in \big[(a_{k-1} + \widehat{w}^{(1)}_k) \wedge a_k, a_k\big)$ or $v \in \big[(b_{\ell-1} + \widehat{w}^{(2)}_\ell) \wedge b_\ell, b_\ell\big)$. This implies that

$$\delta^2(W_{\mathbf{\Theta}^*}, W^*) \leqslant (B-A)^2 \Big(\sum_{k=1}^{K} (w^{(1)}_k - \widehat{w}^{(1)}_k)_+ + \sum_{\ell=1}^{L} (w^{(2)}_\ell - \widehat{w}^{(2)}_\ell)_+ \Big). \tag{2.22}$$

Since U_i are i.i.d. random variables uniformly distributed in $[0,1]$, $n\widehat{w}_k^{(1)}$ follows the binomial distribution $\mathcal{B}(n, w_k^{(1)})$. This implies that $\mathbb{E}[\widehat{w}_k^{(1)}] = w_k^{(1)}$ for every k. Therefore,

$$\begin{aligned}\mathbb{E}[(w_k^{(1)} - \widehat{w}_k^{(1)})_\pm] &= \frac{1}{2}\mathbb{E}[|w_k^{(1)} - \widehat{w}_k^{(1)}|] \\ &\leqslant \Big(\frac{w_k^{(1)}(1-w_k^{(1)})}{4n}\Big)^{1/2} \leqslant \frac{(w_k^{(1)})^{1/2}}{2\sqrt{n}}. \end{aligned} \tag{2.23}$$

Similar upper bound can be obtained for $\mathbb{E}[(w_\ell^{(2)} - \widehat{w}_\ell^{(2)})_+]$. Therefore, combining (2.22), (2.23), the Cauchy-Schwarz inequality and the fact that the weights $w^{(1)}$ sum to one, we get

$$\begin{aligned}\mathbb{E}[\delta^2(W_{\boldsymbol{\Theta}^*}, W^*)] &\leqslant (B-A)^2\Big(\sum_{k=1}^K \mathbb{E}[(w_k^{(1)} - \widehat{w}_k^{(1)})_+] + \sum_{\ell=1}^L \mathbb{E}[(w_\ell^{(2)} - \widehat{w}_\ell^{(2)})_+]\Big) \\ &\leqslant \frac{(B-A)^2}{2}\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big).\end{aligned}$$

We thus conclude the proof by taking the square root of the obtained inequality.

Second claim (Hölder continuous graphon) To make the subsequent formulae more compact, we set $a_n(i) = i/n$ and assume that $\mathcal{L} = 1$. We introduce $I_i = \big[a_n(i-1), a_n(i)\big[$ and $J_j = \big[a_m(j-1), a_m(j)\big[$. For every positive $k \in \mathbb{N}$, let $\mathfrak{S}_n$ be the set of the permutations of $[k]$. For $\sigma \in \mathfrak{S}_k$, let τ_σ be the specific measure-preserving application

$$\tau_\sigma(x) = \sum_{i=1}^k \Big(\frac{\sigma(i)-1}{k} + x - \frac{i-1}{k}\Big)\mathbb{1}_{I_i}(x) \qquad \forall x \in [0,1[, \qquad \tau_\sigma(1) = \frac{\sigma(k)}{k}.$$

Notice that τ_σ corresponds to permutation of intervals $\{I_i : i \in [k]\}$ in accordance with σ.

Using the definition of δ, we have

$$\begin{aligned}\delta^2(W_{\boldsymbol{\Theta}^*}, W^*) &\leqslant \inf_{\substack{\sigma_1\in\mathfrak{S}_n \\ \sigma_2\in\mathfrak{S}_m}} \sum_{i\in[n], j\in[m]} \iint_{I_i\times J_j} \big(\Theta^*_{ij} - W^*(\tau_{\sigma_1}(x), \tau_{\sigma_2}(y))\big)^2 dxdy \\ &= \inf_{\substack{\sigma_1\in\mathfrak{S}_n \\ \sigma_2\in\mathfrak{S}_m}} \sum_{i\in[n], j\in[m]} \iint_{I_{\sigma_1(i)}\times J_{\sigma_2(j)}} \big(\Theta^*_{ij} - W^*(x,y)\big)^2 dxdy\,. \end{aligned} \tag{2.24}$$

Let σ_1 be a random permutation satisfying $U_{(\sigma_1(i))} = U_i$; for example, let σ_1 be such that $U_{\sigma_1^{-1}(1)} \leqslant U_{\sigma_1^{-1}(2)} \leqslant \cdots \leqslant U_{\sigma_1^{-1}(n)}$. We choose σ_2 similarly, so that $V_{(\sigma_2(j))} = V_j$. Recall that $\Theta^*_{ij} = W^*(U_i, V_j)$. Setting $i' = \sigma_1(i)$, $j' = \sigma_2(j)$ and applying the triangle inequality, we get

$$\begin{aligned}|\Theta^*_{i,j} - W^*(x,y)| &= |W^*(x,y) - W^*(U_{(i')}, V_{(j')})| \\ &\leqslant \big|W^*(x,y) - W^*\big(a_{n+1}(i'), a_{m+1}(j')\big)\big| \\ &\quad + \big|W^*\big(a_{n+1}(i'), a_{m+1}(j')\big) - W^*(U_{(i')}, V_{(j')})\big|. \end{aligned} \tag{2.25}$$

If $(x,y) \in I_{\sigma_1(i)} \times J_{\sigma_2(j)} = I_{i'} \times J_{j'}$, as $\big(a_{n+1}(i'), a_{m+1}(j')\big)$ belongs to the same set $I_{i'} \times J_{j'}$, the Hölder property yields

$$\Big(W^*(x,y) - W^*\big(a_{n+1}(i'), a_{m+1}(j')\big)\Big)^2 \leqslant \Big(\frac{1}{n^2} + \frac{1}{m^2}\Big)^{\alpha}. \tag{2.26}$$

For the second term in (2.25), we use again the Hölder property, which leads to

$$\big\{W^*\big(a_{n+1}(i'), a_{m+1}(j')\big) - W^*(U_{(i')}, V_{(j')})\big\}^2 \leqslant \big\{\big|a_{n+1}(i') - U_{(i')}\big|^2 + \big|a_{m+1}(j') - V_{(j')}\big|^2\big\}^{\alpha}.$$

Then, denoting by $\sum_{i,j}$ the double sum $\sum_{i\in[n]} \sum_{j\in[m]}$, we have

$$\begin{aligned}
\mathbb{E}\bigg[\sum_{i,j} \iint_{I_{i'}\times J_{j'}} \Big(\frac{i'}{n+1} - U_{(i')}\Big)^2\bigg] &= \frac{1}{n}\mathbb{E}\Big[\sum_{i=1}^{n}\Big(\frac{i'}{n+1} - U_{(i')}\Big)^2\Big] \\
&= \frac{1}{n}\mathbb{E}\Big[\sum_{i=1}^{n}\Big(\frac{i}{n+1} - U_{(i)}\Big)^2\Big] \\
&\leqslant \max_{i=1,\dots,n}[\mathbf{Var}(U_{(i)})] \leqslant \frac{1}{4n}
\end{aligned} \tag{2.27}$$

where we used the fact that $U_{(i)}$ is drawn from the beta distribution $\beta(i, n+1-i)$. The term with $V_{(j')}$ is treated similarly. Combining (2.24), the Minkowski inequality, (2.26) and (2.27), we get

$$\begin{aligned}
&\mathbb{E}[\delta^2(W_{\mathbf{\Theta}^*}, W^*)]^{1/2} \\
&\quad\leqslant \Big(\frac{1}{n^2} + \frac{1}{m^2}\Big)^{\alpha/2} + \mathbb{E}\bigg[\sum_{i,j} \iint_{I_{i'}\times J_{j'}} \Big(\frac{i'}{n+1} - U_{(i')}\Big)^2 + \Big(\frac{j'}{m+1} - V_{(j')}\Big)^2\bigg]^{\alpha/2} \\
&\quad\leqslant \Big(\frac{1}{n^2} + \frac{1}{m^2}\Big)^{\alpha/2} + \Big(\frac{1}{4n} + \frac{1}{4m}\Big)^{\alpha/2} \leqslant 2\Big(\frac{1}{n} + \frac{1}{m}\Big)^{\alpha/2}.
\end{aligned}$$

This completes the proof.

2.7.3 Proof of Proposition 3 (approximation error for the matrix Θ^*)

Without loss of generality, we prove the desired inequality in the case where $K = \lfloor n/n_0 \rfloor$ and $L = \lfloor m/m_0 \rfloor$. Indeed, if the inequality is true for some value of K, it is necessarily true for any smaller value as well. The same is true for L. Furthermore, we assume $\mathcal{L} = 1$.

We construct the constant-by-block matrix $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$, where $\mathcal{T}$ is defined by (2.3) as follows. Let $n = n_0 K + r$ with $0 \leqslant r < n_0$ and $m = m_0 L + s$ with $0 \leqslant s < m_0$. For all $k \in [K]$ and $\ell \in [L]$, set

$$I_k = [(k-1)n_0 + 1, (kn_0) \wedge n] \qquad \text{and} \qquad J_\ell = [(\ell-1)m_0 + 1, (\ell m_0) \wedge m].$$

The number of integers contained in each set of the form $I_k \times J_\ell$ is denoted by $n_{k,\ell}$. We set

$$\widetilde{Q}_{k,\ell} = \frac{1}{n_{k,\ell}} \sum_{i \in I_k, j \in J_\ell} W^*(U_{(i)}, V_{(j)}).$$

Finally, for every $(i,j) \in [n] \times [m]$, we set

$$\widetilde{\Theta}_{i,j} = \widetilde{Q}_{\lceil \sigma_U(i)/n_0 \rceil, \lceil \sigma_V(j)/m_0 \rceil} = \sum_{k,\ell} \widetilde{Q}_{k,\ell} \mathbb{1}_{\{\sigma_U(i) \in I_k\}} \mathbb{1}_{\{\sigma_V(j) \in J_\ell\}},$$

where σ_U (resp. σ_V) is a permutation of $[n]$ (resp. $[m]$) transforming $(U_i : i \in [n])$ (resp. $(V_j : j \in [m])$) into a nondecreasing sequence. In other terms, $U_{\sigma_U(i)} = U_{(i)}$ and $V_{\sigma_V(j)} = V_{(j)}$ for all i and j. To bound the approximation error we are interested in, note that ($\sum_{k,\ell}$ stands for $\sum_{k \in [K]} \sum_{\ell \in [L]}$)

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{nm} \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\Big] &= \frac{1}{nm} \sum_{k,\ell} \sum_{i : \sigma_U(i) \in I_k} \sum_{j : \sigma_V(j) \in J_\ell} \mathbb{E}\big[(\widetilde{Q}_{k,\ell} - \Theta^*_{ij})^2\big] \\
&= \frac{1}{nm} \sum_{k,\ell} \sum_{i \in I_k} \sum_{j \in J_\ell} \mathbb{E}\Big[\Big(\frac{1}{n_{k,\ell}} \sum_{i' \in I_k} \sum_{j' \in J_\ell} W^*(U_{(i')}, V_{(j')}) - W^*(U_{(i)}, V_{(j)})\Big)^2\Big] \\
\text{(Jensen)} \qquad &\leqslant \frac{1}{nm} \sum_{k,\ell} \sum_{i \in I_k} \sum_{j \in J_\ell} \Big(\frac{1}{n_{k,\ell}} \sum_{i' \in I_k} \sum_{j' \in J_\ell} \mathbb{E}\Big[\big(W^*(U_{(i')}, V_{(j')}) - W^*(U_{(i)}, V_{(j)})\big)^2\Big]\Big).
\end{aligned}$$

Using the Hölder property and the Jensen inequality we obtain

$$\begin{aligned}
\mathbb{E}\Big[\big(W^*(U_{(i')}, V_{(j')}) - W^*(U_{(i)}, V_{(j)})\big)^2\Big] &\leqslant \mathbb{E}\Big[\Big\| \begin{bmatrix} U_{(i')} - U_{(i)} \\ V_{(j')} - V_{(j)} \end{bmatrix} \Big\|^{2\alpha}\Big] \\
&\leqslant \Big\{\mathbb{E}\Big[\Big\| \begin{bmatrix} U_{(i')} - U_{(i)} \\ V_{(j')} - V_{(j)} \end{bmatrix} \Big\|^2\Big]\Big\}^{\alpha}.
\end{aligned}$$

Since $|i - i'| < n_0 + 1 \leqslant (3/2)n_0 \leqslant (3n)/(2K)$, [KTV17, Lemma 4.10, p27] leads to

$$\mathbb{E}\big[|U_{(i')} - U_{(i)}|^2\big] \leqslant 9/(2K)^2.$$

Similarly, $\mathbb{E}\big[|V_{(j')} - U_{(j)}|^2\big] \leqslant 9/(2L)^2$. Therefore,

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{nm} \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\Big]^{1/2} &\leqslant \mathcal{L}\Big[\Big(\frac{3}{2K}\Big)^2 + \Big(\frac{3}{2L}\Big)^2\Big]^{\alpha/2} \\
&\leqslant \frac{3\mathcal{L}}{2}\Big(\frac{1}{K^\alpha} + \frac{1}{L^\alpha}\Big),
\end{aligned}$$

where we have used the inequality $(a+b)^c \leqslant a^c + b^c$ for $a, b \geqslant 0$ and for $c \in (0,1]$.

2.7.4 Proof of Theorem 8 (risk bound for $\widehat{W}^{\mathsf{LS}}$)

First claim: piecewise constant graphon In view of (2.7), the fact that $\widehat{W} = W_{\widehat{\mathbf{\Theta}}}$ and Proposition 2, we have

$$\begin{aligned}\mathbb{E}[\delta(W_{\widehat{\mathbf{\Theta}}}, W^*)^2]^{1/2} &\leqslant \frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}}{\sqrt{nm}} + \mathbb{E}[\delta(W_{\mathbf{\Theta}^*}, W^*)^2]^{1/2} \\ &\leqslant \frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}}{\sqrt{nm}} + \frac{\rho}{\sqrt{2}}\left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\right)^{1/2}. \qquad (2.28)\end{aligned}$$

Let $\widehat{\mathcal{T}}$ and $\mathcal{T}^*$ be the sets of all $n \times m$ matrices with real entries that are constant by block on the same blocks as $\widehat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$, respectively. Clearly, $\widehat{\mathcal{T}}$ and $\mathcal{T}^*$ are linear subspaces of the space of $n \times m$ real matrices equipped with the scalar product $\langle \mathbf{M}_1, \mathbf{M}_2 \rangle = \mathrm{tr}(\mathbf{M}_1^\top \mathbf{M}_2)$. Let $\Pi_{\widehat{\mathcal{T}}}$ be the orthogonal projections onto $\widehat{\mathcal{T}}$. We have $\Pi_{\widehat{\mathcal{T}}}\mathbf{H} = \widehat{\mathbf{\Theta}}$. Therefore,

$$\begin{aligned}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}} &= \|\Pi_{\widehat{\mathcal{T}}}\mathbf{H} - \mathbf{\Theta}^*\|_{\mathsf{F}} \\ &\overset{①}{\leqslant} \|\Pi_{\widehat{\mathcal{T}}}(\mathbf{H} - \mathbf{\Theta}^*)\|_{\mathsf{F}} + \|\Pi_{\widehat{\mathcal{T}}}\mathbf{\Theta}^* - \mathbf{\Theta}^*\|_{\mathsf{F}} \\ &\overset{②}{\leqslant} \|\mathbf{H} - \mathbf{\Theta}^*\|_{\mathsf{F}} + \|(\rho/2)\mathbf{1}_n\mathbf{1}_m^\top - \mathbf{\Theta}^*\|_{\mathsf{F}}.\end{aligned}$$

Above, ① is a consequence of the triangle inequality, whereas ② follows from the fact that $\Pi_{\widehat{\mathcal{T}}}$ is an orthogonal projection (hence, a contraction) and the matrix $(\rho/2)\mathbf{1}_n\mathbf{1}_m^\top$ belongs to the image of $\Pi_{\widehat{\mathcal{T}}}$. Hence

$$\frac{1}{nm}\mathbb{E}\big[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \,\big|\, \boldsymbol{U}, \boldsymbol{V}\big] \leqslant (\sigma + 0.5\rho)^2.$$

For every $k \in [K]$ and $\ell \in [L]$, we define $n_k = n|a_k - a_{k-1}|$, $N_k = \#\{i : U_i \in [a_{k-1}, a_k[\}$, $m_\ell = m|b_\ell - b_{\ell-1}|$ and $M_\ell = \#\{j : V_j \in [b_{\ell-1}, b_\ell[\}$. We also define the event $\Omega_0 = \{N_k \geqslant n_k/2; M_\ell \geqslant m_\ell/2 \text{ for all } k \in [K] \text{ and } \ell \in [L]\}$. Since the event Ω_0^c is $(\boldsymbol{U}, \boldsymbol{V})$-measurable, we get

$$\begin{aligned}\frac{1}{nm}\mathbb{E}\big[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \mathbb{1}_{\Omega_0^c}\big] &= \frac{1}{nm}\mathbb{E}\Big(\mathbb{E}\big[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \,\big|\, \boldsymbol{U}, \boldsymbol{V}\big]\mathbb{1}_{\Omega_0^c}\Big) \\ &\leqslant (\sigma + 0.5\rho)^2\mathbb{P}(\Omega_0^c).\end{aligned}$$

Using the union bound and the Chernoff inequality, one can check that

$$\begin{aligned}\mathbb{P}(\Omega_0^c) &\leqslant \sum_{k=1}^K \mathbb{P}(N_k \leqslant n_k/2) + \sum_{\ell=1}^L \mathbb{P}(M_\ell \leqslant m_\ell/2) \\ &\leqslant \sum_{k=1}^K e^{-n_k/8} + \sum_{\ell=1}^L e^{-m_\ell/8}.\end{aligned}$$

Since we have assumed that $n_k \geqslant 8\log(nK)$ and $m_\ell \geqslant 8\log(mL)$, we get $\mathbb{P}(\Omega_0^c) \leqslant n^{-1} + m^{-1}$. If the parameters n_0 and m_0 used in the definition of the least squares estimator $\widehat{\mathbf{\Theta}}$ satisfy $n_0 = \min_k n_k/2 = n\Delta^{(K)}/2$ and $m_0 = \min_\ell m_\ell/2 = m\Delta^{(L)}/2$, then on the event Ω_0 we ca apply Theorem 6. One can check that $\psi_{n,m}(n_0, m_0) = \psi_{n,m}(\Delta^{(K,L)})$. This, in conjunction with the previous inequalities, implies that

$$\begin{aligned}\frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]}{nm} &= \frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \mathbb{1}_{\Omega_0}] + \mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \mathbb{1}_{\Omega_0^c}]}{nm} \\ &\leqslant \big(25\sigma^2 + 4b\rho\big)\Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big) + \frac{(\sigma + 0.5\rho)^2}{n} + \frac{(\sigma + 0.5\rho)^2}{m} \\ &\leqslant \Big\{\big(27\sigma^2 + 4b\rho\big)^{1/2}\Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} + \frac{\rho}{2}\sqrt{\frac{1}{n} + \frac{1}{m}}\Big\}^2,\end{aligned}$$

under condition that $\psi_{n,m}(\Delta^{(K,L)}) \leqslant (\sigma/b)^2$. One can also check that if $K, L \geqslant 2$ and $n, m \geqslant 5$, it holds

$$\frac{1}{n} + \frac{1}{m} \leq \frac{1}{3}\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big).$$

This inequality, combined with (2.28), completes the proof of the theorem.

Second claim: Hölder continuous graphons Using Equation (2.7) and the Minkowski inequality, we get

$$\begin{aligned}\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} &= \mathbb{E}[\delta(W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}, W^*)^2]^{1/2} \\ &\leqslant \frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}}{\sqrt{nm}} + \mathbb{E}[\delta(W_{\mathbf{\Theta}^*}, W^*)^2]^{1/2}.\end{aligned}$$

Let us set

$$K = L = \Big\lfloor \Big(\frac{3nm\mathcal{L}^2}{25\sigma^2 + 4b\rho}\Big)^{1/2(\alpha+1)} \Big\rfloor.$$

In view of (2.11), we have

$$\begin{aligned}(K/m)^{2(\alpha+1)} &\leqslant \frac{3n\mathcal{L}^2}{(25\sigma^2 + 4b\rho)m^{2\alpha+1}} \\ &\leqslant \frac{3(\sigma/2b)^{4(\alpha+1)} \wedge 1}{\log^4(2n)}. \qquad (2.29)\end{aligned}$$

Let us choose $n_0 = \lfloor n/K \rfloor$ and $m_0 = \lfloor m/K \rfloor$. Thanks to (2.29), we have

$$\frac{m}{K} \geqslant \big(\log^4(2n)\big)^{1/2(\alpha+1)} \geqslant \log 8 > 2.$$

This implies that $m_0 \geqslant 2$ and, therefore $n_0 \geqslant 2$. Using once again (2.29), one can check that

$$\frac{6\log(em/2)}{n_0} \leqslant \frac{6\log(en/2)}{m_0} \leqslant (\sigma/b)^2.$$

As a consequence,

$$\begin{aligned}\psi_{n,m}(n_0, m_0) &= \frac{3\log(en/n_0)}{m_0} + \frac{3\log(em/m_0)}{n_0} \\ &\leqslant \frac{3\log(en/2)}{m_0} + \frac{3\log(em/2)}{n_0} \\ &\leqslant (\sigma/b)^2.\end{aligned}$$

Combining Theorem 6, Proposition 3 and claim 2 of Proposition 2, we arrive at

$$\begin{aligned}\mathbb{E}\big[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2\big]^{1/2} &\leqslant \frac{3\mathcal{L}}{2K^\alpha} + \frac{3\mathcal{L}}{2L^\alpha} + \big(25\sigma^2 + 4b\rho\big)^{1/2}\Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} \\ &\qquad + \frac{2\mathcal{L}}{n^{\alpha/2}} + \frac{2\mathcal{L}}{m^{\alpha/2}} \\ &\leqslant \frac{3\mathcal{L}}{K^\alpha} + \big(25\sigma^2 + 4b\rho\big)^{1/2}\Big(\frac{3K^2}{nm} + \frac{2\log K}{m}\Big)^{1/2} + \frac{4\mathcal{L}}{m^{\alpha/2}} \\ &\leqslant \frac{3\mathcal{L}}{K^\alpha} + 3K\Big(\frac{25\sigma^2 + 4b\rho}{3nm}\Big)^{1/2} + \Big(\frac{(50\sigma^2 + 8b\rho)\log K}{m}\Big)^{1/2} + \frac{4\mathcal{L}}{m^{\alpha/2}}.\end{aligned}$$

In the last display, replacing K with its expression (2.29), we get the claim of the theorem.

2.7.5 Proof of Proposition 4 (relaxation to a linear program)

For further references, we recall that we are interested in solving the problem

$$\min_{\mathbf{Z} \in \mathcal{Z}(n,K,n_0)} \mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}}). \tag{OPT 1}$$

First claim (linearlization of the cost function) We have

$$\begin{aligned}\mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}^{\mathrm{item}}) - \|\mathbf{H}\|_{\mathsf{F}}^2 &= -2\operatorname{tr}(\mathbf{Z}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top \mathbf{H}^\top) + \operatorname{tr}(\mathbf{Z}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top \mathbf{Z}^{\mathrm{item}}\mathbf{Q}^\top \mathbf{Z}^\top) \\ &= -2\operatorname{tr}(\mathbf{Z}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top \mathbf{H}^\top) + \operatorname{tr}(\mathbf{Q}\mathbf{D}\mathbf{Q}^\top \mathbf{Z}^\top \mathbf{Z}).\end{aligned}$$

We will show that $\phi(\mathbf{Z}) = -2\operatorname{tr}(\mathbf{Z}\mathbf{Q}(\mathbf{Z}^{\mathrm{item}})^\top \mathbf{H}^\top) + \operatorname{tr}(\mathbf{Q}\mathbf{D}\mathbf{Q}^\top \mathbf{Z}^\top \mathbf{Z})$. We notice that because of the constraint on the rows of $\mathbf{Z}$, its columns are orthogonal, and $\mathbf{Z}^\top\mathbf{Z} = \mathrm{diag}(n_k; k \in [K])$, where $n_k = \boldsymbol{Z}_{\bullet k}^\top \mathbf{1}_n$ is the number of nonzero entries in the k-th column of $\mathbf{Z}$. So we get

$$\operatorname{tr}(\mathbf{Q}\mathbf{D}\mathbf{Q}^\top \mathbf{Z}^\top \mathbf{Z}) = \sum_{k=1}^K n_k \boldsymbol{Q}_{k,\bullet}\mathbf{D}\boldsymbol{Q}_{k,\bullet}^\top = \sum_{k=1}^K \boldsymbol{Z}_{\bullet k}^\top \mathbf{1}_n \, \boldsymbol{Q}_{k,\bullet}\mathbf{D}\boldsymbol{Q}_{k,\bullet}^\top.$$

This completes the proof of the first claim and entails that (OPT 1) is equivalent to

$$\arg \min_{\mathbf{Z} \in \mathcal{Z}(n,K,n_0)} \phi(\mathbf{Z}). \tag{OPT 2}$$

Second claim (characterization of extreme points) An extreme point of a convex polytope $\mathcal{P}$ is defined as a point in $\mathcal{P}$ that can not be written as a nontrivial convex combination of two elements in $\mathcal{P}$. First, let us prove that any point in $\mathcal{Z}(n, K, n_0)$ is an extreme point. Let $\mathbf{Z} \in \mathcal{Z}(n, K, n_0)$, $\lambda \in (0, 1)$, $\mathbf{Z}_1$ and $\mathbf{Z}_2 \in \widetilde{\mathcal{Z}}(n, K, n_0)$ such that

$$\mathbf{Z} = \lambda \mathbf{Z}_1 + (1 - \lambda)\mathbf{Z}_2 \,.$$

Fix some $i \in [n]$, because of the constrain on the lines of $\mathbf{Z}$, there exist $j \in [K]$ such that

$$1 = Z_{ij} = \lambda (Z_1)_{ij} + (1 - \lambda)(Z_2)_{ij} \,. \tag{2.30}$$

The only way to satisfy (2.30) is to have $(Z_1)_{ij} = (Z_2)_{ij} = 1$ because $\lambda \in (0, 1)$. Then $(Z_1)_{i,\bullet} = (Z_2)_{i,\bullet} = Z_{i,\bullet}$ because of the row constraints. Finally, this holds for each $i \in [n]$, so $\mathbf{Z} = \mathbf{Z}_1 = \mathbf{Z}_2$, which ensures that $\mathbf{Z}$ is an extreme point.

Now it remains to prove that any extreme point of $\widetilde{\mathcal{Z}}(n, K, n_0)$ has all its entries in $\{0, 1\}$. Let $\mathbf{Z} \in \widetilde{\mathcal{Z}}(n, K, n_0)$ which has at least one entry in $(0, 1)$. Let us prove that $\mathbf{Z}$ can not be an extreme point, that is, we can write $\mathbf{Z}$ as a convex combination of two elements in $\widetilde{\mathcal{Z}}(n, K, n_0)$. The proof uses the next two lemmas.

Lemma 3. *Let $i_0 \in [n]$, $k_0 \in [K]$ such that $Z_{i_0 k_0} \in (0, 1)$. Then*

1. *There exists $k_0' \in [K]$ such that $Z_{i_0, k_0'} \in (0, 1)$.*
2. *Either $\sum_{i=1}^n Z_{i,k_0} \notin \mathbb{N}$, or there exists $i_0' \in [n]$ such that $Z_{i_0' k_0} \in (0, 1)$.*

Proof of Lemma 3. The proof is a straightforward consequence of the row and column constraints. Indeed, as the sum of the elements of each line is an integer, if a coefficient is not 0 or 1, then there is another coefficient that lives in $(0, 1)$ on the same line, which proves (i). Moreover, if $\sum_{i=1}^n Z_{i,k_0} \in \mathbb{N}$, the same argument gives that there is another coefficient in $(0, 1)$ on the column k_0. □

If we see matrix $\mathbf{Z}$ as a bi-adjacency matrix of a bipartite graph $\mathcal{G}$, with weighted edges given by the entries of $\mathbf{Z}$, the next lemma formally says that either $\mathcal{G}$ contains a cycle, or it has a path with extreme points that correspond to a column that sums to a number strictly greater than n_0 (see Figure 19).

Lemma 4. *Let $\mathbf{Z} \in \widetilde{\mathcal{Z}}(n, K, n_0) \backslash \{0, 1\}^{n \times K}$. There exists $T \geqslant 1$, and two sequences $(i_t)_{t=1}^T$ and $(k_t)_{t=0}^T$ of different indices (with possibly $k_0 = k_T$) such that*

1. *$Z_{i_t k_t} \in (0, 1)$ for all $1 \leqslant t \leqslant T$.*

2. $Z_{i_{t+1}k_t} \in (0,1)$ *for all* $0 \leqslant t \leqslant T-1$.

3. *Either* $\sum_{i=1}^n Z_{ik_t} > n_0$ *for* $t \in \{0, T\}$ *(we say that* $\mathbf{Z}$ *has a dead end path), or* $k_T = k_0$ *(*$\mathbf{Z}$ *has a cycle).*

Proof of Lemma 4. Let us first assume that all columns of $\mathbf{Z}$ sum to some integers. We denote $Z_{i_1k_0}$ one element of $\mathbf{Z}$ that is in $(0,1)$ According to part (i) of Lemma 3, there exists $k_1 \neq k_0$ such that $Z_{i_1k_1} \in (0,1)$. Following the same Lemma 3, we can find $i_2 \neq i_1$, and $k_2 \neq k_1$ such that $Z_{i_2k_1} \in (0,1)$ and $Z_{i_2k_2} \in (0,1)$ (so $T \geqslant 2$). We iterate the same process until iteration T, with T define as the first time at which we have $i_T = i_{t_0}$ or $k_T = k_{t_0}$ for some $0 \leqslant t_0 \leqslant T-1$, meaning that we met a row or a column we already had in the previous iterations. Notice that $t_0 \leqslant T-2$ because we always have $i_t \neq i_{t+1}$ and $k_t \neq k_{t+1}$. Then we consider the following shifted sequences.

- $i^\# = (i_{t+t_0-1})_{t=1}^{T-t_0}$ and $k_t^\# = k_{t+t_0}$ for $t \in \{0, \dots, T-t_0-1\}$ and $k_{T-t_0}^\# = k_{t_0}$ in the case where $i_T = i_{t_0}$, meaning that we first met a row we already had in the previous iterations. In this case, $t_0 \geqslant 1$.
- $i^\# = (i_{t+t_0})_{t=1}^{T-t_0}$ and $k^\# = (k_{t+t_0})_{t=0}^{T-t_0}$ in the case where $k_T = k_{t_0}$ meaning that we first met a column we already had in the previous iterations.

These sequences $i^\#$ and $k^\#$ satisfy the cycle conditions of the lemma by construction.

Now we assume that there is a column k_0 which has a sum not in $\mathbb{N}$. Then it has a coordinate $Z_{i_1k_0} \in (0,1)$. Applying part (i) of Lemma 3 gives $k_1 \neq k_0$ such that $Z_{i_1k_1} \in (0,1)$.

- If $\sum_{i=1}^n Z_{ik_1} > n_0$, then lemma is proven for $T=1$ and the dead end path setting.
- Else, $\sum_{i=1}^n Z_{ik_1} = n_0$ then we can iterate similarly the previous process until iteration T define as the first time at which we have $i_T = i_{t_0}$ or $k_T = k_{t_0}$ for some $0 \leqslant t_0 \leqslant T-2$ or $\sum_{i=1}^n Z_{ik_T} > n_0$. In the first two cases, we consider the shifted sequences $i^\#$ and $k^\#$ as before, that satisfy the conditions of the lemma. In the last case, the sequences $(i_t)_{t=1}^T$ and $(k_t)_{t=0}^T$ satisfy the desired dead end path conditions.

Notice that $T < +\infty$ because the number of rows and columns is finite. □

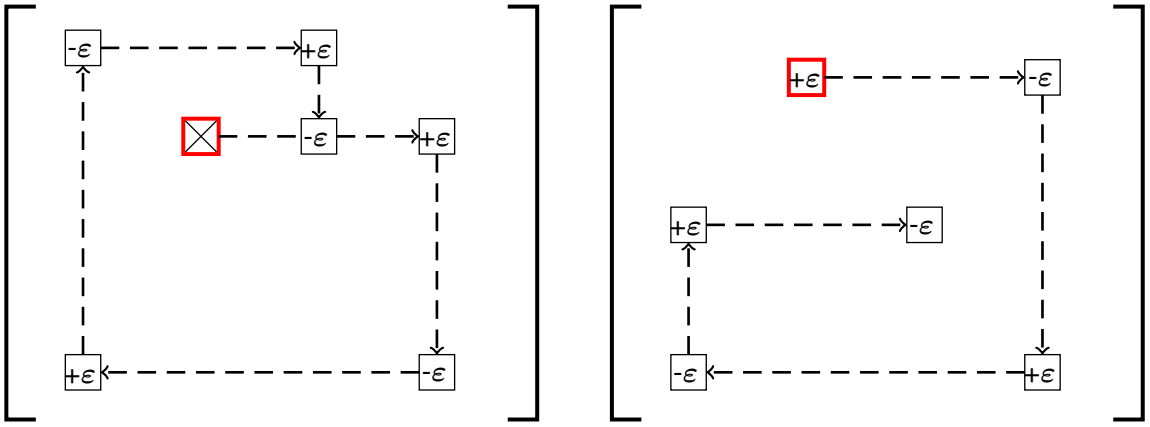
The end of the proof of Proposition 4 is similar to the one in [PC19, Prop 3.4], if we see $\mathbf{Z}$ as the bi-adjacency matrix of a bipartite graph. Let us rewrite the proof adapted for our purpose. We introduce

$$\varepsilon_1 = \min_{k\in[K]} \Big\{ \sum_{i=1}^n \frac{Z_{i,k}-n_0}{2} : \sum_{i=1}^n Z_{i,k} > n_0 \Big\} \wedge 1, \quad \varepsilon_2 = \min_{i,k} \Big\{ \frac{Z_{ik}}{2} : Z_{ik} > 0 \Big\}$$

and $\varepsilon_3 = (1/2)\min_{i,k}\big\{1 - Z_{ik} : Z_{ik} < 1\big\}$. So ε_1, ε_2 and ε_3 are positive real numbers. By convention, the minimum of the empty set is $+\infty$. We finally define $\varepsilon = \min(\varepsilon_1, \varepsilon_2, \varepsilon_3) \in (0,1)$.

Hence, when we add or subtrack ε to one entry of $\mathbf{Z}$ that is in $(0,1)$, it remains in $(0,1)$. Moreover, if this entry is in a column that sums to strictly more than n_0, if we subtrack ε to this entry, the sum of the column remains strictly greater than n_0. We apply Lemma 4 which gives to sequences $(i_t)_{t=1}^T$ and $(k_t)_{t=0}^T$ that satisfy the conditions of the lemma. Then we define $\mathbf{Z}_\varepsilon$ such that (see Figure 19 for a visual construction of $\mathbf{Z}_\varepsilon$)

$$(Z_\varepsilon)_{ij} = \begin{cases} 0 & \text{if} \quad i \neq i_t \text{ or } k \neq k_t \\ +\varepsilon & \text{if} \quad i = i_{t+1}, k = k_t \qquad \text{for some } 0 \leqslant t \leqslant T-1 \\ -\varepsilon & \text{if} \quad i = i_t, k = k_t \qquad \text{for some } 1 \leqslant t \leqslant T\,. \end{cases}$$



(a) Cylcle (b) Dead end path

Figure 19: Examples of matrices $\mathbf{Z}_\varepsilon$: The starting point is represented by a red square. On the left hand side figure we met a line we had before like in the procedure described in Lemma 4. Then we forget the past (what is before this line, that is the crossed-out square), which gives the sequences we need to build $\mathbf{Z}_\varepsilon$. On the right-hand side figure, we never met a line or a column we had before, but we stop because the last column sums to strictly more than n_0, which gives the path to build $\mathbf{Z}_\varepsilon$.

The properties of the sequences imply that

- $\mathbf{Z}_\varepsilon \mathbf{1}_K = \mathbf{0}_n$. Indeed, the i-th row of $\mathbf{Z}$ sums to 0 when $i \notin \{i_1, \ldots, i_T\}$. But if $i = i_t$ for some $t \in [T]$, then there is a $+\varepsilon$ on the k_t-th column, and a $-\varepsilon$ on the k_{t-1}-th column, and 0 anywhere else.
- Moreover we have $\mathbf{Z}_\varepsilon^\top \mathbf{1}_n = \delta_\varepsilon \mathbf{1}_K$ where $\delta_\varepsilon \in \{0, \pm\varepsilon\}$. Indeed, the k-th column sums to 0 for $k \notin \{k_0, \ldots, k_T\}$. For $k = k_t$ with $t \in [T-1]$, then there is a $+\varepsilon$ on the i_{t+1}-th row and a $-\varepsilon$ on the i_t-th row and 0 anywhere else. For $t \in \{0, T\}$ we have two cases. If $k_T = k_0$, then the k_0-th column sums to 0. If $k_T \neq k_0$, there is only a $+\varepsilon$ on the k_0-th column and only a $-\varepsilon$ on the k_T-th column.

The choice of ε ensures that $\mathbf{Z}_+ = \mathbf{Z} + \mathbf{Z}_\varepsilon$ and $\mathbf{Z}_- = \mathbf{Z} - \mathbf{Z}_\varepsilon$ live in $\widetilde{\mathcal{Z}}(n, K, n_0)$. Finally, $\mathbf{Z} = (\mathbf{Z}_+ + \mathbf{Z}_-)/2$ and $\mathbf{Z}$ can not be an extreme point.

2.7.6 Proof of Theorem 9 (lower bounds)

Although the general structure of the proof and some important parts of it are similar to those of the proof of [KTV17, Proposition 3.4], there are some technical differences that are due to the fact that the graphon and the observed matrix are not symmetric. Furthermore, the bounds involve some terms depending on m and m, which was not the case in lower bounds proved in the literature.

To get the desired lower bound, we divide the problem into the following three minimax lower bounds

$$\inf_{\widehat{W}} \sup_{W^* \in \mathcal{W}_\rho[K,L]} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] \geqslant c\rho^2\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big) \tag{2.31}$$

$$\inf_{\widehat{W}} \sup_{W^* \in \mathcal{W}_\rho[K,L]} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] \geqslant c\rho\Big(\frac{KL}{Nnm} \wedge \rho\Big) \tag{2.32}$$

$$\inf_{\widehat{W}} \sup_{W^* \in \mathcal{W}_\rho[2,2]} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] \geqslant c\rho\Big(\frac{1}{N\sqrt{nm}} \wedge \rho\Big) \tag{2.33}$$

If these three inequalities hold true, then the desired result will be true with the constant $c/3$. The rest of this section is split into three subsections, each of which contains the proof of one of the inequalities (2.31), (2.32) and (2.33).

Proof of (2.31): error due to the unknown partition

Since (K, n) and (L, m) play symmetric roles, we will only prove the lower bound $\rho^2\sqrt{K/n}$. The same arguments will lead to the lower bound $\rho^2\sqrt{L/m}$. As a consequence, we will get the lower bound $\rho^2(\sqrt{K/n} \vee \sqrt{L/m}) \geqslant \rho^2(\sqrt{K/n} + \sqrt{L/m})/2$.

Without loss of generality, we can assume that K is a multiple of 16. Indeed, by choosing $C > 17$, for any $K > 17$, setting $K' = 16\lfloor K/16 \rfloor$, the inequality

$$\inf_{\widehat{W}} \sup_{W^* \in \mathcal{W}_\rho[K',L]} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] \geqslant c\rho^2\sqrt{\frac{K'}{n}}$$

would imply

$$\inf_{\widehat{W}} \sup_{W^* \in \mathcal{W}_\rho[K,L]} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] \geqslant \inf_{\widehat{W}} \sup_{W^* \in \mathcal{W}_\rho[K',L]} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] \geqslant \frac{c}{4}\rho^2\sqrt{\frac{K}{n}}.$$

To establish the desired lower bound, we follow the standard recipe [Tsy08, Theorem 2.7] consisting in designing a finite set of graphons that has the following two properties: the graphons from this set are well separated when the distance is measured by the metric δ^2 and, in the same time, the distributions generated by these graphons are close, which makes it difficult to differentiate them based on the data matrix $\mathbf{H}$.

To define this set, we choose a $K \times L$ matrix $\mathbf{Q}$ with entries from $\{0, \rho\}$ and two positive numbers $\varepsilon, \varepsilon'$; conditions on $\mathbf{Q}$, ε and ε' will be specified later. For any $K \in \mathbb{N}$, define

$$\mathscr{C}^K = \Big\{ \boldsymbol{s} \in \{-1,+1\}^K : \sum_{k=1}^K s_k = 0 \Big\}, \quad I^a_{\boldsymbol{s},\varepsilon} = \Big[\sum_{k=1}^{a-1} \Big(\frac{1}{K} + \varepsilon s_k \Big), \sum_{k=1}^{a} \Big(\frac{1}{K} + \varepsilon s_k \Big) \Big)$$

for $a \in [K]$ and $\boldsymbol{s} \in \mathscr{C}^K$, with the convention that $\sum_{k=1}^0 = 0$. Similarly, for every $\boldsymbol{t} \in \mathscr{C}^L$, we set $J^b_{\boldsymbol{t},\varepsilon'} = \big[\sum_{\ell=1}^{b-1} \big(\frac{1}{L} + \varepsilon' t_\ell \big), \sum_{\ell=1}^{b} \big(\frac{1}{L} + \varepsilon' t_\ell \big) \big)$. The length of an interval I will be denoted by $|I|$. We can now define the class of graphons $W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}$ for each $\boldsymbol{s} \in \mathscr{C}^K$ and $\boldsymbol{t} \in \mathscr{C}^L$ by

$$W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(u,v) = \sum_{k=1}^K \sum_{\ell=1}^L Q_{k,\ell} \mathbb{1}_{I^k_{\boldsymbol{s},\varepsilon}}(u) \mathbb{1}_{J^\ell_{\boldsymbol{t},\varepsilon'}}(v).$$

We denote by $\mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}}$ the distribution of $\mathbf{H} = (H_{ij}, i \in [n], j \in [m])$, where $\mathbf{H}$ is sampled according to the Binomial model with graphon $W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}$. The next three lemmas, the proofs of which are postponed to the end of this subsection, provide the main technical tools necessary to establish the desired lower bound.

Lemma 5. *If $4K\varepsilon \leqslant 1$, then for all $\boldsymbol{s}$ and $\boldsymbol{s}'$ from $\mathscr{C}^K$ and all $\boldsymbol{t} \in \mathscr{C}^L$, the following inequality $D_{\mathsf{KL}}(\mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}} || \mathbf{P}_{W_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}}) \leqslant 6n(K\varepsilon)^2$ holds true.*

Lemma 6. *Assume that K, L are large enough integers multiple of 16 and satisfying $KL \geqslant L\log^2 L + K \log^2 K$. There exists $\mathbf{B} \in \{-1,1\}^{K \times L}$ satisfying the following two properties.*

i) For all $(k_1, k_2) \in [K]^2$, $k_1 \neq k_2$, and for all $(\ell_1, \ell_2) \in [L]^2$, $\ell_1 \neq \ell_2$, it holds that

$$|\langle \mathbf{B}_{k_1 \bullet}, \mathbf{B}_{k_2 \bullet} \rangle| \leqslant L/4 \qquad \text{and} \qquad |\langle \mathbf{B}_{\bullet \ell_1}, \mathbf{B}_{\bullet \ell_2} \rangle| \leqslant K/4$$

ii) Let $\pi_i : [K/16] \to [K]$ and $\nu_i : [L/16] \to [L]$, $i = 1, 2$ be arbitrary bijections such that either $\mathrm{Im}(\pi_1) \cap \mathrm{Im}(\pi_2) = \varnothing$ or $\mathrm{Im}(\nu_1) \cap \mathrm{Im}(\nu_2) = \varnothing$. Then

$$\sum_{k=1}^{K/16} \sum_{\ell=1}^{L/16} \big(B_{\pi_1(k),\nu_1(\ell)} - B_{\pi_2(k),\nu_2(\ell)} \big)^2 \geqslant \frac{KL}{512}.$$

Let us denote for each $\boldsymbol{s} \in \mathscr{C}^K$, $\mathcal{A}_{\boldsymbol{s}} = \{k \in [K] : s_k = 1\}$. Notice that $|\mathcal{A}_{\boldsymbol{s}}| = K/2$.

Lemma 7. *Choose $\mathbf{Q} = \mathbf{B} + \mathbf{1}_K \mathbf{1}_L^\top$, where $\mathbf{B}$ is given by Lemma 6. Let $\boldsymbol{s}$ and $\boldsymbol{s}'$ be two distinct vectors from $\mathscr{C}^K$ such that $|\mathcal{A}_{\boldsymbol{s}} \Delta \mathcal{A}_{\boldsymbol{s}'}| \geqslant K/4$. For $\varepsilon' = 1/(4L)$, for any $K \in \mathbb{N}$, $\varepsilon \in [0, 1/(4K)]$ and $\boldsymbol{t} \in \mathscr{C}^L$, we have*

$$\delta^2(W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}, W_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}) \geqslant \frac{\varepsilon K \rho^2}{512}.$$

Lemma 4.4 in [KTV17] implies that there exists a subset $\widetilde{\mathscr{C}}^K \subset \mathscr{C}^K$ such that $\log |\widetilde{\mathscr{C}}^K| \geqslant K/16$ and $|\mathcal{A}_{\boldsymbol{s}} \triangle \mathcal{A}_{\boldsymbol{s}'}| \geqslant K/4$ for any $\boldsymbol{s} \neq \boldsymbol{s}'$ from $\widetilde{\mathscr{C}}^K$. We consider the set $\{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon} : \boldsymbol{s} \in \widetilde{\mathscr{C}}^K\}$ for

a fixed $\boldsymbol{t} \in \mathscr{C}^L$ and for $\varepsilon^{-1} = 2^4\sqrt{6nK}$. In view of Lemma 5 and Lemma 7, for any $\boldsymbol{s}, \boldsymbol{s}' \in \widetilde{\mathscr{C}}^K$ such that $\boldsymbol{s} \neq \boldsymbol{s}'$, we have

$$\delta^2(W_{\boldsymbol{s},\varepsilon}, W_{\boldsymbol{s}',\varepsilon}) \geqslant \frac{\rho^2}{2^{13}}\sqrt{\frac{K}{6n}} \quad \text{and} \quad D_{\mathsf{KL}}(\mathbf{P}_{W_{\boldsymbol{s},\varepsilon}} || \mathbf{P}_{W_{\boldsymbol{s}',\varepsilon}}) \leqslant \frac{1}{16^2}K \leqslant \frac{1}{16}\log|\widetilde{\mathscr{C}}^K|.$$

Therefore, we can apply [Tsy08, Theorem 2.7] to get

$$\inf_{\widehat{W}} \sup_{W^* \in \mathcal{W}_\rho[K,L]} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] \geqslant c\rho^2\sqrt{\frac{K}{n}}$$

for some universal constant $c > 0$. This completes the proof of (2.31), module the proofs of three technical lemmas appended below.

Proof of Lemma 5. One can check that, for every matrix $\mathbf{A} \in [N]^{n\times m}$, and for every $\boldsymbol{s} \in \mathscr{C}^K$,

$$\mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}}(\mathbf{H} = \mathbf{A}) = \prod_{i=1}^n\prod_{j=1}^m \int_{[0,1]^2} \binom{N}{A_{ij}} W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(u_i, v_j)^{A_{ij}}(1 - W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(u_i, v_j))^{N-A_{ij}}\, du_i\, dv_j.$$

Using the fact that $W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}$ is piecewise constant, we get

$$\begin{aligned}
\mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}}(\mathbf{H} = \mathbf{A}) &= \prod_{i,j}\sum_{k=1}^K\sum_{\ell=1}^L \int_{I^k_{\boldsymbol{s},\varepsilon}}\int_{J^\ell_{\boldsymbol{t},\varepsilon'}} \binom{N}{A_{ij}} Q_{k\ell}^{A_{i,j}}(1 - Q_{k\ell})^{N-A_{i,j}}\, du_i\, dv_j \\
&= \prod_{i,j}\sum_{k,\ell} \binom{N}{A_{ij}} Q_{k\ell}^{A_{i,j}}(1 - Q_{k\ell})^{N-A_{i,j}} \big|I^k_{\boldsymbol{s},\varepsilon}\big| \cdot \big|J^\ell_{\boldsymbol{t},\varepsilon'}\big| \\
&= \sum_{\mathbf{K},\mathbf{L}}\prod_{i,j} \binom{N}{A_{ij}} Q_{K_{ij}L_{ij}}^{A_{ij}}(1 - Q_{K_{ij}L_{ij}})^{N-A_{ij}} \big|I^{K_{ij}}_{\boldsymbol{s},\varepsilon}\big| \cdot \big|J^{L_{ij}}_{\boldsymbol{t},\varepsilon'}\big|,
\end{aligned}$$

where the outer sum of the last line is over all matrices $\mathbf{K}$ and $\mathbf{L}$ having entries respectively in $[K]$ and in $[L]$. Let us define

$$\begin{aligned}
\Psi(\mathbf{K}, \mathbf{L}, \mathbf{A}) &= \prod_{i,j} \binom{N}{A_{ij}} Q_{K_{ij}L_{ij}}^{A_{ij}}(1 - Q_{K_{ij}L_{ij}})^{N-A_{ij}}, \\
w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K}, \mathbf{L}) &= \prod_{i,j} \big|I^{K_{ij}}_{\boldsymbol{s},\varepsilon}\big| \cdot \big|J^{L_{ij}}_{\boldsymbol{t},\varepsilon'}\big|.
\end{aligned}$$

The computations above imply that

$$\begin{aligned}
D_{\mathsf{KL}}(\mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}} || \mathbf{P}_{W_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}}) &= \sum_{\mathbf{A}\in\{0,1\}^{n\times m}} \mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}}(\mathbf{H} = \mathbf{A}) \log\Big(\frac{\mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}}(\mathbf{H} = \mathbf{A})}{\mathbf{P}_{W_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}}(\mathbf{H} = \mathbf{A})}\Big) \\
&= \sum_{\mathbf{A},\mathbf{K},\mathbf{L}} \Psi(\mathbf{K}, \mathbf{L}, \mathbf{A}) w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K}, \mathbf{L}) \log\Big(\frac{\sum_{\mathbf{K}',\mathbf{L}'}\Psi(\mathbf{K}', \mathbf{L}', \mathbf{A}) w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K}', \mathbf{L}')}{\sum_{\mathbf{K}',\mathbf{L}'}\Psi(\mathbf{K}', \mathbf{L}', \mathbf{A}) w_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}(\mathbf{K}', \mathbf{L}')}\Big).
\end{aligned}$$

It is clear that $w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K}, \mathbf{L}) \geqslant 0$ and $\sum_{\mathbf{K},\mathbf{L}} w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K}, \mathbf{L}) = 1$. Since the function $(x, y) \mapsto x\log(x/y)$

is convex, we apply the Jensen inequality to get

$$D_{\mathsf{KL}}(\mathbf{P}_{W_{\boldsymbol{s},\boldsymbol{t}\varepsilon}}||\,\mathbf{P}_{W_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}}) \leqslant \sum_{\mathbf{A},\mathbf{K},\mathbf{L}} \Psi(\mathbf{K},\mathbf{L},\mathbf{A}) w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L}) \log\Big(\frac{\Psi(\mathbf{K},\mathbf{L},\mathbf{A}) w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}{\Psi(\mathbf{K},\mathbf{L},\mathbf{A}) w_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}\Big)$$
$$= \sum_{\boldsymbol{K},\mathbf{L}} w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L}) \log\Big(\frac{w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}{w_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}\Big).$$

The last expression can be seen as the Kullback-Leibler divergence between two product distributions on $[K]^{n\times m}\times[L]^{n\times m}$. Since the Kullback-Leibler divergence between product distributions is the sum of Kullback-Leibler divergences, we get

$$\sum_{\mathbf{K},\mathbf{L}} w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L}) \log\Big(\frac{w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}{w_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}\Big) = n\sum_{k=1}^{K} |I^k_{\boldsymbol{s},\varepsilon}| \log\Big(\frac{|I^k_{\boldsymbol{s},\varepsilon}|}{|I^k_{\boldsymbol{s}',\varepsilon}|}\Big)$$
$$\leqslant n\sum_{k=1}^{K} \frac{\big(|I^k_{\boldsymbol{s},\varepsilon}| - |I^k_{\boldsymbol{s}',\varepsilon}|\big)^2}{|I^k_{\boldsymbol{s}',\varepsilon}|},$$

where the last inequality follows from the fact that the Kullback-Leibler divergence does not exceed the chi-square divergence. Since $|I^k_{\boldsymbol{s}',\varepsilon}| = (1/K) + \varepsilon s'_k \geqslant (1/K) - \varepsilon \geqslant 3/(4K)$, we get

$$\sum_{\mathbf{K},\mathbf{L}} w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L}) \log\Big(\frac{w_{\boldsymbol{s},\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}{w_{\boldsymbol{s}',\boldsymbol{t},\varepsilon}(\mathbf{K},\mathbf{L})}\Big) \leqslant n\sum_{k=1}^{K} \frac{\big(\varepsilon s_k - \varepsilon s'_k\big)^2}{3/(4K)}$$
$$\leqslant 6nK^2\varepsilon^2.$$

This completes the proof of the lemma. □

Proof of Lemma 6. Let $\boldsymbol{\Xi}$ be a $K\times L$ random matrix with iid Rademacher entries $\xi_{k,\ell}$, *i.e*, $\mathbf{P}(\xi_{k,\ell}=\pm 1)=1/2$. Then $\langle \boldsymbol{\Xi}_{k_1\bullet}, \boldsymbol{\Xi}_{k_2\bullet}\rangle = \sum_{\ell=1}^{L}\xi_{k_1,\ell}\xi_{k_2,\ell}$. By the Hoeffding inequality

$$\mathbf{P}(|\langle \boldsymbol{\Xi}_{k_1\bullet}, \boldsymbol{\Xi}_{k_2\bullet}\rangle| \geqslant L/4) \leqslant 2e^{-L/32}\,.$$

By the union bound, we obtain that for all $k_1\neq k_2\in[K]$, $|\langle \boldsymbol{\Xi}_{k_1\bullet}, \boldsymbol{\Xi}_{k_2\bullet}\rangle| \leqslant L/4$ with probability larger than $1-2K^2e^{-L/32}$, which is larger than $3/4$ for $L\geqslant 480$. Similarly, one checks that

$$\mathbf{P}(\max_{\ell_1\neq\ell_2} |\langle \boldsymbol{\Xi}_{\bullet,\ell_1}, \boldsymbol{\Xi}_{\bullet,\ell_2}\rangle| \leqslant K/4) > 3/4.$$

Thus, we get

$$\mathbf{P}\Big(\max_{\ell_1\neq\ell_2} |\langle \boldsymbol{\Xi}_{\bullet,\ell_1}, \boldsymbol{\Xi}_{\bullet,\ell_2}\rangle| \leqslant K/4 \text{ and } \max_{k_1\neq k_2} |\langle \boldsymbol{\Xi}_{k_1,\bullet}, \boldsymbol{\Xi}_{k_2,\bullet}\rangle| \leqslant L/4\Big) > 1/2. \tag{2.34}$$

For the second property stated in the lemma, we fix some $\mathcal{X}_i, \mathcal{Y}_i$ and π_i, ν_i as in the statement

and define

$$T(\pi_{1:2},\nu_{1:2},\Xi)=\frac{1}{4}\sum_{k=1}^{K/16}\sum_{\ell=1}^{L/16}\big(\xi_{\pi_1(k),\nu_1(\ell)}-\xi_{\pi_2(k),\nu_2(\ell)}\big)^2.$$

Clearly, $T[\pi_{1:2},\nu_{1:2},\Xi]$ is a sum of $KL/2^8$ i.i.d Bernouilli random variables with parameter 1/2. Applying again the Hoeffding inequality, we have

$$\mathbf{P}\Big(T[\pi_{1:2},\nu_{1:2},\Xi]\leqslant\frac{KL}{2^{11}}\Big)=\mathbf{P}\Big(\frac{KL}{2^9}-T\geqslant\frac{3KL}{2^{11}}\Big)\leqslant e^{-9KL/2^{13}}.$$

There are no more than $(K/16)!^2(L/16)!^2$ functions π_1,π_2,ν_1,ν_2 satisfying conditions of ii). Therefore, the union bound implies that with probability at least $1-(K/16)!^2(L/16)!^2e^{-9KL/2^{13}}$, we have $T[\pi_{1:2},\nu_{1:2},\Xi]\geqslant KL/2^{11}$ for all π_1,π_2,ν_1,ν_2. Choosing K and L large enough, and using the condition $KL\geqslant K\log^2K+L\log^2L$, we get that

$$\mathbf{P}\Big(\min_{\pi_1,\pi_2,\nu_1,\nu_2}\sum_{k=1}^{K/16}\sum_{\ell=1}^{L/16}\big(\xi_{\pi_1(k),\nu_1(\ell)}-\xi_{\pi_2(k),c\nu_2(\ell)}\big)^2\geqslant\frac{KL}{512}\Big)>1/2. \tag{2.35}$$

Combining (2.34) and (2.35), we get that the probability that the random matrix Ξ satisfies properties i) and ii) is strictly positive. This implies that the set of such matrices is not empty. □

Proof of Lemma 7. Without loss of generality, throughout this proof, we assume that $\rho=1$. Furthermore, since $\boldsymbol{t}$ is fixed, we will often drop it in the notation and write $W_{\boldsymbol{s},\varepsilon}$ instead of $W_{\boldsymbol{s},\boldsymbol{t},\varepsilon}$.

It suffices to prove that for all measure preserving bijections $\tau_1:[0,1]\to[0,1]$ and $\tau_2:[0,1]\to[0,1]$,

$$\|W_{\boldsymbol{s},\varepsilon}-W_{\boldsymbol{s}',\varepsilon}\circ(\tau_1\otimes\tau_2)\|_{\mathbb{L}_2}^2\geqslant\frac{\varepsilon K}{512}.$$

If $u\in I^k_{\boldsymbol{s},\varepsilon}$ and $u'\in I^{k'}_{\boldsymbol{s},\varepsilon}$ for some $k,k'\in[K]$, then

$$\begin{aligned}\Big|\int_0^1(W_{\boldsymbol{s},\varepsilon}(u,v)-1/2)(W_{\boldsymbol{s},\varepsilon}(u',v)-1/2)dv\Big|&=\Big|\sum_{\ell=1}^{L}(1/L+\varepsilon' t_\ell)\Big(Q_{k,\ell}-\frac{1}{2}\Big)\Big(Q_{k',\ell}-\frac{1}{2}\Big)\Big|\\&\leqslant\frac{1}{4L}|\langle\mathbf{B}_{k,\bullet},\mathbf{B}_{k',\bullet}\rangle|+\frac{L\varepsilon'}{4}\leqslant\frac{1}{8}.\end{aligned} \tag{2.36}$$

For $k,k'\in[K]$, let $\omega_{kk'}=\lambda\big\{I^k_{\boldsymbol{s},\varepsilon}\cap\tau_1^{-1}\big(I^{k'}_{\boldsymbol{s}',\varepsilon}\big)\big\}$ where λ is the Lebesgue measure on $\mathbb{R}$. Notice that $\sum_{k=1}^K\omega_{kk'}=(1/K)+\varepsilon s'_{k'}$ and $\sum_{k'=1}^K\omega_{kk'}=(1/K)+\varepsilon s_k$. We also introduce $h_{\boldsymbol{s},k}(v)=W_{\boldsymbol{s},\varepsilon}(u^k_{\boldsymbol{s},\varepsilon},v)-1/2$, where $u^k_{\boldsymbol{s},\varepsilon}$ is any point from $I^k_{\boldsymbol{s},\varepsilon}$. We have

$$\|W_{\boldsymbol{s},\varepsilon}-W_{\boldsymbol{s}',\varepsilon}(\tau_1\otimes\tau_2)\|_{\mathbb{L}_2}^2=\sum_{k=1}^K\sum_{k'=1}^K\omega_{k,k'}\|h_{\boldsymbol{s},k}-h_{\boldsymbol{s}',k'}\circ\tau_2\|_{\mathbb{L}_2}^2.$$

In view of the fact that $|h_{\boldsymbol{s}',k'}(v)| = 1/2$ for all $v \in [0,1]$ and (2.36), for any $k' \neq k$, we have

$$\begin{aligned}\|h_{\boldsymbol{s}',k'} \circ \tau_2 - h_{\boldsymbol{s}',k''} \circ \tau_2\|^2_{\mathbb{L}_2} &= \|h_{\boldsymbol{s}',k'}\|^2_{\mathbb{L}_2} + \|h_{\boldsymbol{s}',k''}\|^2_{\mathbb{L}_2} - 2\langle h_{\boldsymbol{s}',k'}, h_{\boldsymbol{s}',k''}\rangle \\ &\geqslant 1/2 - 1/4 = 1/4.\end{aligned}$$

By the triangle inequality

$$\|h_{\boldsymbol{s},k} - h_{\boldsymbol{s}',k'} \circ \tau_2\|_{\mathbb{L}_2} + \|h_{\boldsymbol{s},k} - h_{\boldsymbol{s}',k''} \circ \tau_2\|_{\mathbb{L}_2} \geqslant \|h_{\boldsymbol{s}',k'} \circ \tau_2 - h_{\boldsymbol{s}',k''} \circ \tau_2\|_{\mathbb{L}_2} \geqslant 1/2\,.$$

As a consequence, for any $k \in [K]$, there exists at most one $k' \in [K]$ such that $\|h_{\boldsymbol{s},k} - h_{\boldsymbol{s}',k'} \circ \tau_2\|_{\mathbb{L}_2} < 1/4$. If such a k' exists, we denote it by $\pi(k)$. If it does not exist, we set $\pi(k) = k$. Using the same arguments, for any $k' \in [K]$, there is at most one $k \in [K]$ such that $\|h_{\boldsymbol{s},k} - h_{\boldsymbol{s}',k'} \circ \tau_2\|_{\mathbb{L}_2} < 1/4$. This implies that π is injective and then it is a permutation of $[K]$. Furthermore, we get

$$\begin{aligned}\|W_{\boldsymbol{s},\varepsilon} - W_{\boldsymbol{s}',\varepsilon} \circ (\tau_1 \otimes \tau_2)\|^2_{\mathbb{L}_2} &\geqslant \frac{1}{16}\sum_{k=1}^{K}\sum_{k' \neq \pi(k)} \omega_{k,k'} \\ &= \frac{1}{16}\sum_{k=1}^{K}(1/K + \varepsilon s_k - \omega_{k,\pi(k)})\,. \qquad (2.37)\end{aligned}$$

If the sum $\sum_{k=1}^{K}(1/K + \varepsilon s_k - \omega_{k,\pi(k)})$ is larger than $K\varepsilon/16$, then the lemma is proved.

In the sequel, we check that the same is true if $\sum_{k=1}^{K}(1/K + \varepsilon s_k - \omega_{k,\pi(k)}) < K\varepsilon/16$ as well. Note that the last inequality can be rewritten as $\sum_{k=1}^{K}\omega_{k,\pi(k)} > 1 - K\varepsilon/16$. Let us show that the cardinality of the set $A = \{k \in \mathcal{A}_{\boldsymbol{s}} : s'_{\pi(k)} > 0 \text{ and } \omega_{k,\pi(k)} \geqslant 1/K\}$ is at least $7K/16$. Indeed, notice that because $\omega_{k,k'} \leqslant \big(1/K + \varepsilon s_k\big) \wedge \big(1/K + \varepsilon s_{k'}\big)$, $\omega_{k,\pi(k)} \geqslant 1/K$ implies $s_k > 0$ and $s'_{\pi(k)} > 0$. Therefore,

$$\begin{aligned}1 - \frac{K\varepsilon}{16} \leqslant \sum_{k=1}^{K}\omega_{k,\pi(k)} &= \sum_{k:s_k<0}\omega_{k,\pi(k)} + \sum_{k\in A}\omega_{k,\pi(k)} + \sum_{\substack{k\notin A\\ s_k>0}}\omega_{k,\pi(k)} \\ &\leqslant \frac{K}{2}\Big(\frac{1}{K} - \varepsilon\Big) + |A|\Big(\frac{1}{K} + \varepsilon\Big) + \Big(\frac{K}{2} - |A|\Big)\frac{1}{K} = 1 + \varepsilon\Big(|A| - \frac{K}{2}\Big),\end{aligned}$$

which leads to $|A| \geqslant 7K/16$.

Since $\boldsymbol{s}, \boldsymbol{s}' \in \widetilde{\mathscr{C}}^K$ are such that $|\mathcal{A}_{\boldsymbol{s}} \triangle \mathcal{A}_{\boldsymbol{s}'}| \geqslant K/4$, we have $|\mathcal{A}_{\boldsymbol{s}} \cap \mathcal{A}_{\boldsymbol{s}'}| \leqslant 3K/8$. Let us choose $B \subset A \cap \mathcal{A}^c_{\boldsymbol{s}'}$ of cardinality $K/16$ and set $C = \pi(B)$. Such a choice is possible since

$$\begin{aligned}|A \cap \mathcal{A}^c_{\boldsymbol{s}'}| = |A| - |A \cap \mathcal{A}_{\boldsymbol{s}'}| &\geqslant \frac{7K}{16} - |\mathcal{A}_{\boldsymbol{s}} \cap \mathcal{A}_{\boldsymbol{s}'}| \\ &\geqslant \frac{7K}{16} - \frac{3K}{8} = \frac{K}{16}.\end{aligned}$$

Note also that $B \cap C = \varnothing$. Indeed, if $k \in B$, then $\pi(k) \in \mathcal{A}_{\boldsymbol{s}'}$. Therefore, $\pi(k) \notin B$ meaning

that $k \notin C$.

For $\ell, \ell' \in [L]$, let $\omega'_{\ell,\ell'} = \lambda\big\{J^{\ell}_{\boldsymbol{t},\varepsilon'} \cap \tau_2^{-1}(J^{\ell'}_{\boldsymbol{t},\varepsilon'})\big\}$. Using the same arguments as above, we obtain the existence of a permutation $\nu : [L] \to [L]$ such that, akin to (2.37),

$$\|W_{\boldsymbol{s},\varepsilon} - W_{\boldsymbol{s}',\varepsilon} \circ (\tau_1 \otimes \tau_2)\|^2_{\mathbb{L}_2} \geqslant \frac{1}{16} \sum_{b=1}^{L} (1/L + \varepsilon' t_\ell - \omega'_{\ell,\nu(\ell)}).$$

Define the set $A' = \{\ell \in [L] : t_\ell > 0, \omega'_{\ell,\nu(\ell)} \geqslant 1/L\}$. If $|A'| \leqslant L/16$, then

$$\begin{aligned}
\sum_{\ell=1}^{L} (1/L + \varepsilon' t_\ell - \omega'_{\ell,\nu(\ell)}) &\geqslant 1 - \sum_{t_\ell<0} \omega'_{\ell,\nu(\ell)} - \sum_{t_\ell>0} \omega'_{\ell,\nu(\ell)} (\mathbb{1}_{\ell \notin A'} + \mathbb{1}_{\ell \in A'}) \\
&\geqslant 1 - \frac{L}{2}\Big(\frac{1}{L} - \varepsilon'\Big) - \Big(\frac{L}{2} - |A'|\Big)\frac{1}{L} - |A'|\Big(\frac{1}{L} + \varepsilon'\Big) = \Big(\frac{L}{2} - |A'|\Big)\varepsilon' \\
&\geqslant \frac{7L\varepsilon'}{16} \geqslant 1/16
\end{aligned}$$

and therefore $\|W_{\boldsymbol{s},\varepsilon} - W_{\boldsymbol{s}',\varepsilon} \circ (\tau_1 \otimes \tau_2)\|^2_{\mathbb{L}_2} \geqslant 1/256 \geqslant K\varepsilon/64$, where we used that $4K\varepsilon < 1$.

Suppose now that $|A'| > L/16$. Let B' be an arbitrary subset of A' of cardinality $L/16$. We have

$$\begin{aligned}
\|W_{\boldsymbol{s},\varepsilon} - W_{\boldsymbol{s}',\varepsilon} \circ (\tau_1 \otimes \tau_2)\|^2_{\mathbb{L}_2} &\geqslant \sum_{k \in B} \sum_{\ell \in B'} \int_{I^k_{\boldsymbol{s},\varepsilon} \times J^{\ell}_{\boldsymbol{s}',\varepsilon}} \big(W_{\boldsymbol{s},\varepsilon} - W_{\boldsymbol{s}',\varepsilon}(\tau_1 \otimes \tau_2)\big)^2(u,v)\, du\, dv \\
&\geqslant \sum_{k \in B} \sum_{\ell \in B'} \omega_{k,\pi(k)} \omega'_{\ell,\nu(\ell)} (Q_{k,\ell} - Q_{\pi(k),\nu(\ell)})^2 \\
&\geqslant \frac{1}{4KL} \sum_{k \in B} \sum_{\ell \in B'} (B_{k,\ell} - B_{\pi(k),\nu(\ell)})^2.
\end{aligned}$$

By Lemma 6, the last term is larger than $1/(4 \times 512) \geqslant K\varepsilon/512$, and the claim of the lemma follows. □

Proof of (2.32): error due to the unknown values of the graphon

Similarly to the previous proof, we will use [Tsy08, Theorem 2.7], which needs a class of graphons that are well separated for the distance δ and that generate similar distributions on the space of $n \times m$ matrices. In this proof, all the graphons of the set will have the same partitions and will differ only by the values of the function taken on this partition.

Let $\mathcal{Q}_0 = \{(\rho/2)(1-\varepsilon), (\rho/2)(1+\varepsilon)\}^{K \times L}$ be the set of all $K \times L$ matrices with entries equal either $(\rho/2)(1-\varepsilon)$ or $(\rho/2)(1+\varepsilon)$, where $\varepsilon \in (0, 1/2)$ will be specified later. For any $\mathbf{Q} \in \mathcal{Q}_0$, we define the graphon

$$W_{\mathbf{Q}}(u,v) = \sum_{k=1}^{K} \sum_{\ell=1}^{L} Q_{k,\ell} \mathbb{1}_{[(k-1)/K, k/K)}(u) \mathbb{1}_{[(\ell-1)/L, \ell/L)}(v)\,.$$

We need two technical lemmas for completing the proof. These lemmas are stated below, whereas their proofs are postponed to the end of this subsection. For any pair of permutations $\pi : [K] \to [K]$ and $\nu : [L] \to [L]$, and any matrix $\mathbf{Q}$, we denote by $\mathbf{Q}^{\pi,\nu}$ the matrix with permuted rows and columns $Q^{\pi,\nu}_{k,\ell} = Q_{\pi(k),\nu(\ell)}$.

Lemma 8. *For K and L large enough satisfying $KL \geqslant L\log^2 L + K\log^2 K$, there exists a set $\mathcal{Q} \subset \mathcal{Q}_0$ satisfying $\log|\mathcal{Q}| \geqslant KL/32$ and $\min_{\pi,\nu} \|\mathbf{Q}_1 - \mathbf{Q}_2^{\pi,\nu}\|_{\mathsf{F}}^2 \geqslant \rho^2\varepsilon^2 KL/8$ for every $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathcal{Q}$ such that $\mathbf{Q}_1 \neq \mathbf{Q}_2$.*

Lemma 9. *The following assertions hold true*

1. *If $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are such that $\min_{\pi,\nu} \|\mathbf{Q}_1 - \mathbf{Q}_2^{\pi,\nu}\|_{\mathsf{F}}^2 \geqslant \rho^2\varepsilon^2 KL/8$, then $\delta^2(W_{\mathbf{Q}_1}, W_{\mathbf{Q}_2}) \geqslant \rho^2\varepsilon^2/8$.*
2. *For any pair of matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ from $\mathcal{Q}$, we have $D_{\mathsf{KL}}(\mathbf{P}_{W_{\mathbf{Q}_1}} || \mathbf{P}_{W_{\mathbf{Q}_2}}) \leqslant 6Nnm\rho\varepsilon^2$.*

We set $\varepsilon^2 = \frac{1}{54\times 32}\big(\frac{KL}{Nnm\rho} \wedge 1\big)$, which allows us to apply [Tsy08, Theorem 2.7], since in view of Lemma 9 and Lemma 8,

$$D_{\mathsf{KL}}(\mathbf{P}_{W_{\mathbf{Q}_1}} || \mathbf{P}_{W_{\mathbf{Q}_2}}) \leqslant \frac{1}{9}\log|\mathcal{Q}|, \qquad \delta^2(W_{\mathbf{Q}_1}, W_{\mathbf{Q}_2}) \geqslant \frac{1}{13824}\Big(\frac{\rho KL}{Nnm} \wedge \rho^2\Big), \qquad \forall\, \mathbf{Q}_1, \mathbf{Q}_2 \in \mathcal{Q}.$$

This completes the proof of (2.32).

Proof of Lemma 8. Without loss of generality, we assume in this proof that $\rho = 2$. We define the pseudo-distance $\delta(\mathbf{Q}_1, \mathbf{Q}_2) = \min_{\pi,\nu} \|\mathbf{Q}_1 - \mathbf{Q}_2^{\pi,\nu}\|_{\mathsf{F}}$, where the minimum is taken over all the permutations of $[K]$ and $[L]$. Let $\mathcal{Q}$ be a maximal subset of $\mathcal{Q}$ of matrices $\mathbf{Q}$ that are $r := \rho\varepsilon\sqrt{KL/2}$-separated with respect to δ. By maximality of $\mathcal{Q}$, we have the inclusion

$$\mathcal{Q} \subset \bigcup_{\mathbf{Q}\in\mathcal{Q}} \mathbb{B}_\delta(\mathbf{Q}, r),$$

where $\mathbb{B}_\delta(\mathbf{Q}, r)$ is the ball centered at $\mathbf{Q}$ with radius r with respect to δ. So $|\mathcal{Q}| \cdot |\mathbb{B}_\delta(\mathbf{Q}, r)| \geqslant 2^{KL}$ since all the balls have a the same cardinality. Notice that $\mathbb{B}_\delta(\mathbf{Q}, r) \subset \bigcup_{\pi,\nu} \mathbb{B}_{\mathsf{F}}(\mathbf{Q}^{\pi,\nu}, r)$ yielding $|\mathbb{B}_\delta(\mathbf{Q}, r)| \leqslant K!L!\,|\mathbb{B}_{\mathsf{F}}(\mathbf{Q}, r)|$. If $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathcal{Q}$, we have $\|\mathbf{Q}_1 - \mathbf{Q}_2\|_{\mathsf{F}}^2 = 4\rho^2\varepsilon^2 d_{\mathsf{H}}(\mathbf{Q}_1, \mathbf{Q}_2)$ with d_{H} the Hamming distance. Then $\mathbb{B}_{\mathsf{F}}(\mathbf{Q}, r) = \mathbb{B}_{\mathsf{H}}(\mathbf{Q}, r^2/(2\rho\varepsilon)^2)$ with $r^2/(2\rho\varepsilon)^2 = KL/8$. The Varshamov-Gilbert lemma [Tsy08, Lemma 2.9] yields

$$|\mathcal{Q}'| \cdot |\mathbb{B}_{\mathsf{H}}(\mathbf{Q}, KL/8)| \leqslant 2^{KL}$$

with $\mathcal{Q}'$ a maximal subset of matrices $\mathbf{Q}$ that are $KL/8$-separated, and $|\mathcal{Q}'| \geqslant 2^{KL/8}$. Thus we obtain

$$\begin{aligned} |\mathcal{Q}| &\geqslant \frac{2^{KL}}{|\mathbb{B}_\delta(\mathbf{Q}, r)|} \geqslant \frac{2^{KL}}{K!L!\,|\mathbb{B}_{\mathsf{F}}(\mathbf{Q}, r)|} \\ &= \frac{2^{KL}}{K!L!\,|\mathbb{B}_{\mathsf{H}}(\mathbf{Q}, KL/8)|} \geqslant \frac{2^{KL}}{K!L!2^{KL}2^{-KL/8}} \end{aligned}$$

$$= \exp\Big(\frac{KL}{8}\log 2 - \log(K!L!)\Big)$$
$$\geqslant \exp\Big(\frac{KL}{16} - K\log K - L\log L\Big).$$

The last term is larger than $e^{KL/32}$ for K and L greater than some constants. □

Proof of Lemma 9. **First claim** Let $\tau_1, \tau_2 : [0,1] \to [0,1]$ be two measure preserving bijections. We want to prove that

$$\|W_{\mathbf{Q}_1} - W_{\mathbf{Q}_2} \circ (\tau_1 \otimes \tau_2)\|_{\square}^2 \geqslant \varepsilon^2/8.$$

For any $k, k' \in [K]$, let $\omega_{k,k'} = \lambda\big([(k-1)/K, k/K] \cap \tau_1^{-1}\big([(k'-1)/K, k'/K]\big)\big)$ where λ is the Lebesgue measure on $[0,1]$. Similarly, for any $\ell, \ell' \in [L]$, let $\omega'_{\ell,\ell'} = \lambda\big([(\ell-1)/L, \ell/L] \cap \tau_2^{-1}\big([(\ell'-1)/L, \ell'/L]\big)\big)$.

We have that $\sum_k \omega_{k,k'} = 1/K = \sum_{k'} \omega_{k,k'}$, that is the matrix $M\omega$ is doubly stochastic. For any permutation π of $[K]$, denote $A(\pi)$ the corresponding permutation matrix. By the Birkhoff-von Neumann theorem, $M\omega$ is a convex combination of permutation matrices, so there exist positive numbers γ_π such that $\omega = \sum_\pi \gamma_\pi A(\pi)$ and $\sum_\pi \gamma_\pi = 1/K$, where the sums are taken over all the permutations of $[K]$. Thus

$$\begin{aligned}
\|W_{\mathbf{Q}_1} - W_{\mathbf{Q}_2} \circ (\tau_1 \otimes \tau_2)\|_{\square}^2 &= \sum_{k,k' \in [K]} \sum_{\ell,\ell' \in [L]} \omega_{k,k'}\omega'_{\ell,\ell'} |(Q_1)_{k,\ell} - (Q_2)_{k',\ell'}|^2 \\
&= \sum_{\pi,\nu} \sum_{k,k' \in [K]} \sum_{\ell,\ell' \in [L]} \gamma_\pi \gamma_\nu A(\pi)_{k,k'} A(\nu)_{\ell,\ell'} |(Q_1)_{k,\ell} - (Q_2)_{k',\ell'}|^2 \\
&= \sum_{\pi,\nu} \sum_{k \in [K]} \sum_{\ell \in [L]} \gamma_\pi \gamma_\nu |(Q_1)_{k,\ell} - (Q_2)_{\pi(k),\nu(\ell)}|^2 \\
&= \sum_{\pi,\nu} \gamma_\pi \gamma_\nu \underbrace{\|\mathbf{Q}_1 - \mathbf{Q}_2^{\pi,\nu}\|_{\mathsf{F}}^2}_{\geqslant \rho^2\varepsilon^2 KL/8} \geqslant \frac{\rho^2\varepsilon^2}{8}
\end{aligned}$$

and the claim of the lemma follows.

Second claim Let $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ be a vector of n i.i.d random variables uniformly distributed on $[K]$. We also denote by $\boldsymbol{\chi} = (\chi_1, \dots, \chi_m)$ a vector of m i.i.d random variables uniformly distributed on $[L]$. Let $\boldsymbol{\Theta}_1 \in [0,1]^{n \times m}$ with entries $(\Theta_1)_{ij} = (\mathbf{Q}_1)_{\zeta_i,\chi_j}$. Assume that $\mathbf{H}_1$ is, conditionally on $\boldsymbol{\zeta}$ and $\boldsymbol{\chi}$, a matrix sampled according to the binomial model with parameter N and the probability matrix $\boldsymbol{\Theta}_1$. Notice that $\mathbf{H}_1$ has distribution $\mathbf{P}_{W_{\mathbf{Q}_1}}$. We introduce $\alpha_{\mathbf{a}} = \mathbf{P}(\boldsymbol{\zeta} = \mathbf{a})$, $\beta_{\mathbf{b}} = \mathbf{P}(\boldsymbol{\chi} = \mathbf{b})$ and $p^{(1)}_{\mathbf{Hab}} = \mathbf{P}(\mathbf{H}_1 = \mathbf{H} | \boldsymbol{\zeta} = \mathbf{a}, \boldsymbol{\xi} = \mathbf{b})$ for any $\mathbf{a} \in [K]^n, \mathbf{b} \in [L]^m$ and $\mathbf{H} \in \{0,1\}^{n \times m}$. We have similar notation $p^{(2)}_{\mathbf{Hab}}$ replacing the indices above. Then

$$D_{\mathsf{KL}}(\mathbf{P}_{W_{\mathbf{Q}_1}}, \mathbf{P}_{W_{\mathbf{Q}_2}}) = \sum_{\mathbf{H}} \sum_{\mathbf{a}} \sum_{\mathbf{b}} \alpha_{\mathbf{a}}\beta_{\mathbf{b}} p^{(1)}_{\mathbf{Hab}} \log\Big(\frac{\sum_{\mathbf{a}}\sum_{\mathbf{b}} \alpha_{\mathbf{a}}\beta_{\mathbf{b}} p^{(1)}_{\mathbf{Hab}}}{\sum_{\mathbf{a}}\sum_{\mathbf{b}} \alpha_{\mathbf{a}}\beta_{\mathbf{b}} p^{(2)}_{\mathbf{Hab}}}\Big)$$

$$\text{(Jensen)} \qquad \leqslant \sum_{\mathbf{a}} \sum_{\mathbf{b}} \alpha_{\mathbf{a}} \beta_{\mathbf{b}} \sum_{\mathbf{H}} p^{(1)}_{\mathbf{Hab}} \log \left(\frac{p^{(1)}_{\mathbf{Hab}}}{p^{(2)}_{\mathbf{Hab}}} \right).$$

When $\mathbf{a}$ and $\mathbf{b}$ are fixed, the sum over $\mathbf{H}$ is the Kullback-Leibler divergence between two nm-product of Binomial measures, each of which has as parameter either (N,p) or (N,q) with $p := \rho(1+\varepsilon)/2$ and $q := \rho(1-\varepsilon)/2$. This gives

$$D_{\mathsf{KL}}(\mathbf{P}_{W_{\mathbf{Q}_1}} || \, \mathbf{P}_{W_{\mathbf{Q}_2}}) \leqslant Nnm\kappa(p,q)$$

where $\kappa(p,q)$ is the Kullback-Leibler divergence between two Bernoulli measures with parameter p and q respectively. We have $\kappa(p,q) \leqslant (p-q)^2(p^{-1}+q^{-1}) = 4\rho\varepsilon^2/(1-\varepsilon^2) \leqslant 16\rho\varepsilon^2/3$. This completes the proof of the lemma. □

Proof of (2.33)

Fix some $\varepsilon \in (0, 1/4)$, and let $W_1 \equiv \rho/2$ be a constant graphon. We define also $W_2(u,v) = \rho(1/2+\varepsilon)$ if $(u,v) \in [0,1/2)^2 \cup [1/2,1]^2$ and $W_2(u,v) = \rho(1/2-\varepsilon)$ elsewhere. We get

$$\delta(W_1, W_2) = \rho\varepsilon \,.$$

Thus we have

$$\begin{aligned}
\inf_{\widehat{W}} \max_{W^* \in \{W_1, W_2\}} \mathbb{E}_{W^*}\big[\delta^2(\widehat{W}, W^*)\big] &\geqslant \frac{1}{2}\Big(\int \delta^2(\widehat{W}, W_1)\, d\mathbf{P}_{W_1} + \int \delta^2(\widehat{W}, W_2)\, d\mathbf{P}_{W_2} \Big) \\
&\geqslant \frac{1}{2} \int \delta^2(\widehat{W}, W_1) + \delta^2(\widehat{W}, W_2)\, \min(d\mathbf{P}_{W_1}, d\mathbf{P}_{W_2}) \\
&\geqslant \frac{\delta^2(W_2, W_1)}{4} \int \min(d\mathbf{P}_{W_1}, d\mathbf{P}_{W_2}) \\
&\geqslant \frac{\rho^2\varepsilon^2}{8} \exp(-\chi^2(\mathbf{P}_{W_1} || \, \mathbf{P}_{W_2})),
\end{aligned}$$

where $\chi^2(\mathbf{P}_{W_1} || \, \mathbf{P}_{W_2})$ stands for the chi-square divergence between $\mathbf{P}_{W_1}$ and $\mathbf{P}_{W_1}$. In the last inequality, we used (2.24) and (2.26) from [Tsy08]. Finally, the next lemma gives an upper-bound on the chi-square divergence, which allows us to complete the proof, taking $\varepsilon^2 = c_0/12\rho N\sqrt{nm}$.

Lemma 10. *There exists an absolute constant $c_0 > 0$ such that $\chi^2(\mathbf{P}_{W_1} || \, \mathbf{P}_{W_2}) \leqslant 1/4$ provided that $12\rho N\sqrt{nm}\varepsilon^2 \leqslant c_0$.*

Proof. Let $L(\mathbf{H})$ be the Radon-Nikodym derivative of $\mathbf{P}_{W_2}$ with respect to $\mathbf{P}_{W_1}$. We have that $\chi^2(\mathbf{P}_{W_1} || \, \mathbf{P}_{W_2}) = \mathbb{E}_{W_1}[L(\mathbf{H})^2] - 1$, so it remains to prove that $\mathbb{E}_{W_1}[L(\mathbf{H})^2] \leqslant 5/4$. In the sequel, $\mathbb{E}[\,\cdot\,]$ will refer to $\mathbb{E}_{W_1}[\,\cdot\,]$. We also introduce $p_0 = \rho/2$, $p_1 = \rho(1/2+\varepsilon)$ and $p_2 = \rho(1/2-\varepsilon)$.

As the graphon $W_2 \in \mathcal{W}_\rho[2,2]$, we can assume that $\{U_i\}$ are i.i.d Bernouilli random variables with parameter 1/2, and similarly for $\{V_j\}$. Given $\{U_i\}$ and $\{V_j\}$, define the set

$S = \{(a,b) : U_a = V_b\}$. For $(i,j) \in S$ (resp. S^c), H_{ij} has Binomial distribution of parameter (N, p_1) (resp. (N, p_2)). Let μ be the distribution of S, then we have

$$L(\mathbf{H}) = \int L_S(\mathbf{H})\, d\mu(S)$$

with

$$L_S(\mathbf{H}) = \Big(\frac{1-p_1}{1-p_0}\Big)^{N|S|}\Big(\frac{1-p_2}{1-p_0}\Big)^{Nnm-N|S|}\prod_{(a,b)\in S}\Big(\frac{p_1(1-p_0)}{p_0(1-p_1)}\Big)^{H_{a,b}}\prod_{(a,b)\in S^c}\Big(\frac{p_2(1-p_0)}{p_0(1-p_2)}\Big)^{H_{a,b}}.$$

By Fubini theorem, we can write $\mathbb{E}[L(\mathbf{H})^2] = \int \mathbb{E}[L_{S_1}(\mathbf{H})L_{S_2}(\mathbf{H})]\, d\mu(S_1)d\mu(S_2)$ with

$$\begin{aligned}\mathbb{E}[L_{S_1}(\mathbf{H})L_{S_2}(\mathbf{H})] =& \\ &\Big(\frac{1-p_1}{1-p_0}\Big)^{N(|S_1|+|S_2|)}\Big(\frac{1-p_2}{1-p_0}\Big)^{N(2nm-|S_1|-|S_2|)}\mathbb{E}\Bigg[\prod_{(a,b)\in S_1\cap S_2}\Big(\frac{p_1(1-p_0)}{p_0(1-p_1)}\Big)^{2H_{a,b}}\\ &\prod_{(a,b)\in S_1^c\cap S_2^c}\Big(\frac{p_2(1-p_0)}{p_0(1-p_2)}\Big)^{2H_{a,b}}\prod_{(a,b)\in S_1\Delta S_2}\Big(\frac{p_1p_2(1-p_0)^2}{p_0^2(1-p_1)(1-p_2)}\Big)^{H_{a,b}}\Bigg].\end{aligned}$$

Recall that expectation is taken with respect to $\mathbf{P}_{W_1}$. We can also use the independence of variables H_{ij} conditionally on $\{U_i\}, \{V_j\}$ to get

$$\begin{aligned}\mathbb{E}[L_{S_1}(\mathbf{H})L_{S_2}(\mathbf{H})] =& \Big[1+\frac{(p_1-p_0)^2}{p_0(1-p_0)}\Big]^{N|S_1\cap S_2|}\Big[1+\frac{(p_2-p_0)^2}{p_0(1-p_0)}\Big]^{N|S_1^c\cap S_2^c|}\\ &\times\Big[1+\frac{p_1p_2+p_0^2-p_1p_0-p_2p_0}{p_0(1-p_0)}\Big]^{N|S_1\Delta S_2|}\\ =&\Big[1+\frac{\rho^2\varepsilon^2}{p_0(1-p_0)}\Big]^{N|S_1\cap S_2|+N|S_1^c\cap S_2^c|}\Big[1-\frac{\rho^2\varepsilon^2}{p_0(1-p_0)}\Big]^{N|S_1\Delta S_2|}\\ \leqslant&\Big[1+\frac{\rho^2\varepsilon^2}{p_0(1-p_0)}\Big]^{N|S_1\cap S_2|+N|S_1^c\cap S_2^c|-N|S_1\Delta S_2|}\\ \leqslant&\exp\Big[\Big({}^3\!/_2|S_1\cap S_2|+{}^3\!/_2|S_1^c\cap S_2^c|-nm\Big)4N\rho\varepsilon^2\Big].\end{aligned}$$

Thus, it remains to bound an exponential moment of $T = |S_1 \cap S_2| + |S_1^c \cap S_2^c|$ where S_1 and S_2 are independent and distributed according to μ. We denote by $\{U_i\}$ and $\{U_i'\}$ the variables that aim to generate S_1 and S_2 respectively, and similarly for $\{V_j\}$ and $\{V_j'\}$. For $(i,j) \in \{0,1\}^2$, define

$$N_{ij} = |\{a, U_a = i \text{ and } U_a' = j\}| \quad \text{and} \quad M_{ij} = |\{a, V_a = i \text{ and } V_a' = j\}.$$

Then we get

$$|S_1 \cap S_2| = N_{00}M_{00} + N_{01}M_{01} + N_{10}M_{10} + N_{11}M_{11} \quad \text{and}$$

$$|S_1^c \cap S_2^c| = N_{00}M_{11} + N_{01}M_{10} + N_{10}M_{01} + N_{11}M_{00}$$

which implies that $T = (N_{00}+N_{11})(M_{00}+M_{11})+(N_{01}+N_{10})(M_{01}+M_{10})$. Notice that $N_{00}+N_{11}$ follows a Binomial distribution with parameters $(n, 1/2)$ and $N_{01} + N_{10} = n - (N_{00} + N_{11})$ (and similarly replacing N by M). Define $X = (N_{00} + N_{11}) - n/2$ and $Y = (M_{00} + M_{11}) - m/2$, we have

$$\begin{aligned} T &= (X + n/2)(Y + m/2) + (n/2 - X)(m/2 - Y) \\ &= 2XY + \frac{nm}{2}\,. \end{aligned}$$

This gives

$$\begin{aligned} \mathbb{E}[L^2(H)] &\leqslant \mathbb{E}_{(X,Y)}\Big[\exp\Big((XY - \frac{nm}{4})12N\rho\varepsilon^2\Big)\Big] \\ &\leqslant \mathbb{E}_{(X,Y)}\Big[\exp\Big(12N\rho\varepsilon^2 XY\Big)\Big]\,. \end{aligned}$$

As X (resp. Y) is sub-Gaussian, with sub-Gaussian norm less than $\sqrt{n}$ (resp. $\sqrt{m}$), XY is sub-exponential with sub-exponential norm upper bounded by $\sqrt{nm}$. This entails that there exists a constant c_0 such that $\mathbb{E}\big[\exp\big(12N\rho\varepsilon^2 XY\big)\big] \leqslant 5/4$ as soon as $12N\rho\varepsilon^2\sqrt{nm} \leqslant c_0$. □

The three lower-bounds have been proved, which completes the proof of Theorem 9.

2.8 Auxiliary results

Lemma 11. *Let X be a random variable and $a \in \mathbb{R}$, $b, c, d \geqslant 0$ be some constants. If*

$$\mathbb{P}(X \geqslant a + bt + ct^2) \leqslant de^{-t} \qquad \textit{for all } t \geqslant 0,$$

then $\mathbb{E}[X] \leqslant a + bd + 2cd$.

Proof. In the case $c = 0$, this inequality is well-known. Therefore, we consider only the case $c > 0$. Without loss of generality, we assume that $a = 0$ and $c = 1$. Indeed, we can always reduce to this case by considering the random variable $X' = (X - a)_+/c$ with $b' = b/c$. Thus, we know that $\mathbb{P}(X \geqslant t^2 + bt) \leqslant de^{-t}$ for every $t \geqslant 0$. Note that the condition $b \geqslant 0$ entails that the mapping $t \mapsto t^2 + bt$ defined on $[0, +\infty)$ is bijective. Setting $z = t^2 + bt$, this implies that

$$\mathbb{P}\big(X \geqslant z\big) \leqslant d\exp\big\{(b/2) - \sqrt{z + (b/2)^2}\big\}, \qquad \forall z \geq 0.$$

This inequality yields

$$\mathbb{E}[X] \leqslant d\int_0^\infty \exp\big\{(b/2) - \sqrt{z + (b/2)^2}\big\}\, dz$$

$$= d \int_0^\infty e^{-t}(2t + b)\, dt$$
$$= bd + 2d.$$

This completes the proof. □

Definition 1. We say that a zero-mean random variable ζ satisfies the (a, b)-Bernstein condition, if we have

$$\mathbb{E}[e^{\lambda\zeta}] \leqslant \exp\left\{\frac{\lambda^2 a}{2(1 - b|\lambda|)}\right\} \qquad \text{provided that} \quad |\lambda| \leqslant 1/b.$$

One can show that if ξ satisfies the (a, b)-Bernstein condition, then the variance of ξ is bounded from above by a. Indeed, since $x^2 \leqslant 2(e^x - 1 - x - x^3/6)$ for every $x \in \mathbb{R}$, replacing x by $\lambda\xi$ for λ small enough and taking the expectation, we get

$$\begin{aligned}
\lambda^2\mathbb{E}[\xi^2] &\leqslant 2(\mathbb{E}[e^{\lambda\xi}] - 1 - \lambda\mathbb{E}[\xi] - (\lambda^3/6)\mathbb{E}[\xi^3]) \\
&\leqslant 2\Big(\exp\left\{\frac{\lambda^2 a}{2(1 - b|\lambda|)}\right\} - 1\Big) - (\lambda^3/3)\mathbb{E}[\xi^3] \\
&= \frac{\lambda^2 a}{1 - b|\lambda|} + o(\lambda^2)
\end{aligned}$$

as $\lambda \to 0$. Dividing the two sides of the last inequality by λ^2 and letting λ go to zero, we get $\mathbf{Var}[\xi] = \mathbb{E}[\xi^2] \leqslant a$.

Lemma 12 (Bernstein inequality)**.** *Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ be a zero-mean random vector with independent coordinates and let $\boldsymbol{\alpha} \in \mathbb{R}^n$ be a deterministic vector. Assume that for some $a, b > 0$, all ξ_i's satisfy the (a, b)-Bernstein condition. Then, for every $\delta \in (0, 1)$, we have*

$$\mathbb{P}\Big(\boldsymbol{\alpha}^\top\boldsymbol{\xi} \leqslant \sqrt{2a\log(1/\delta)}\,\|\boldsymbol{\alpha}\|_2 + b\|\boldsymbol{\alpha}\|_\infty \log(1/\delta)\Big) \geqslant 1 - \delta,$$
$$\mathbb{P}\Big(|\boldsymbol{\alpha}^\top\boldsymbol{\xi}| \leqslant \sqrt{2a\log(2/\delta)}\,\|\boldsymbol{\alpha}\|_2 + b\|\boldsymbol{\alpha}\|_\infty \log(2/\delta)\Big) \geqslant 1 - \delta.$$

Proof. Without loss of generality, we assume that $\|\boldsymbol{\alpha}\|_\infty = 1$. The Markov inequality yields

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\alpha}^\top\boldsymbol{\xi} \geqslant t) &\leqslant e^{-\lambda t}\mathbb{E}[e^{\lambda\boldsymbol{\alpha}^\top\boldsymbol{\xi}}] = e^{-\lambda t}\prod_{i\in[n]}\mathbb{E}[e^{\lambda\alpha_i\xi_i}] \\
&\leqslant \exp\Big(-\lambda t + \frac{\lambda^2 a\|\boldsymbol{\alpha}\|_2^2}{2(1 - b|\lambda|)}\Big), \qquad \forall\lambda : |\lambda| \leqslant 1/b.
\end{aligned}$$

One can check that, for $h(u) = 1 + u - \sqrt{1 + 2u}$, $u > 0$,

$$\sup_{|\lambda|\leqslant 1/b}\Big(\lambda t - \frac{\lambda^2 A}{2(1 - b|\lambda|)}\Big) = \frac{A}{b^2}h\Big(\frac{bt}{A}\Big) \geqslant \frac{t^2}{2(A + bt)}, \qquad \forall t > 0.$$

Hence, we get

$$\mathbb{P}(\boldsymbol{\alpha}^\top \boldsymbol{\xi} \geqslant t) \leqslant \exp\Big\{ - \frac{a\|\boldsymbol{\alpha}\|_2^2}{b^2} h\Big(\frac{bt}{a\|\boldsymbol{\alpha}\|_2^2}\Big)\Big\}, \qquad \forall t \in \mathbb{R}.$$

Thus, if $bt/a\|\boldsymbol{\alpha}\|_2^2 \geqslant h^{-1}(b^2 z/a\|\boldsymbol{\alpha}\|_2^2)$, we have $\mathbb{P}(\boldsymbol{\alpha}^\top \boldsymbol{\xi} \geqslant t) \leqslant e^{-z}$. One can check that $h^{-1}(v) = v + \sqrt{2v}$, which implies that the above condition of t is equivalent to

$$\begin{aligned} t &\geqslant \frac{a\|\boldsymbol{\alpha}\|_2^2}{b}\Big(\frac{b^2 z}{a\|\boldsymbol{\alpha}\|_2^2} + \frac{b\sqrt{2z}}{\sqrt{a}\|\boldsymbol{\alpha}\|_2}\Big) \\ &= bz + \|\boldsymbol{\alpha}\|_2\sqrt{2az}. \end{aligned}$$

Replacing z by $\log(1/\delta)$, we get the first inequality of the lemma.

For the second inequality, we simply remark that the moment generating function of the random variables $-\xi_i$ satisfy the same assumption as the one of ξ_i's. Therefore, we have

$$\begin{aligned} \mathbb{P}\Big(\boldsymbol{\alpha}^\top \boldsymbol{\xi} \leqslant \sqrt{2a\log(1/\delta)}\,\|\boldsymbol{\alpha}\|_2 + b\log(1/\delta)\Big) &\geqslant 1-\delta, \\ \mathbb{P}\Big(-\boldsymbol{\alpha}^\top \boldsymbol{\xi} \leqslant \sqrt{2a\log(1/\delta)}\,\|\boldsymbol{\alpha}\|_2 + b\log(1/\delta)\Big) &\geqslant 1-\delta. \end{aligned}$$

Using the union bound and replacing δ by $\delta/2$, we get the second claim of the lemma. □

Lemma 13. *Let $B > 0$ and $N \in \mathbb{N}$.*

1. *If ζ is a zero mean random variable such that $\mathbb{P}(|\zeta| \leqslant B) = 1$, then it satisfies the (a,b)-Bernstein condition with $a = \mathbf{Var}[\zeta]$ and $b = B/3$.*
2. *If ζ is a the average of N independent zero-mean random variables each of which takes values in $[-B, B]$, then it satisfies the (a,b)-Bernstein condition with $a = \mathbf{Var}[\zeta]$ and $b = B/(3N)$.*
3. *If $\zeta = (\zeta' - \mathbb{E}[\zeta'])/N$ with ζ' drawn from a Poisson distribution with intensity $N\theta$, then it satisfies the (a,b)-Bernstein condition with $a = \mathbf{Var}[\zeta]$ and $b = 1/(3N)$.*

Proof. The proof of these claims is quite standard and based on the inequality $e^z - z - 1 \leqslant 3z^2/(3-|z|)$ provided that $|z| < 3$. □

Lemma 14. *Let $\xi_1, \dots, \xi_N$ be centered and independent random variables satisfying (σ_ξ^2, b_ξ)-Bernstein condition. For every partition $G = \{A_1, \dots, A_k\}$ of $[N]$, let us define the projection matrix Π_G by*

$$(\Pi_G \boldsymbol{v})_i = \frac{1}{|A_j|}\sum_{\ell \in A_j} v_\ell \quad \textit{if} \quad i \in A_j, \quad \forall \boldsymbol{v} \in \mathbb{R}^N.$$

We have that with probability at least $1 - e^{-t}$ *for all* $t > 0$

$$\forall \boldsymbol{v},\ \boldsymbol{v}^\top \Pi_G \boldsymbol{\xi} \leqslant \frac{1}{2}\|\Pi_G \boldsymbol{\xi}\|_2 \|\boldsymbol{v}\|_2 + \sigma_\xi \sqrt{2(t + \log M)}\|\boldsymbol{v}\|_2 + b_\xi(t + \log M)\|\boldsymbol{v}\|_\infty.$$

where $\log M \leqslant k \log 12 \leqslant 2.5k$. *On the same event, we have*

$$\|\Pi_G \boldsymbol{\xi}\|_2^2 \leqslant 4(t + \log M)\big(2\sigma_\xi^2 + b_\xi \|\Pi_G \boldsymbol{\xi}\|_\infty\big).$$

Proof. Let $\boldsymbol{w} \in \mathbb{R}^N$ be a vector such that $\|\pi_G \boldsymbol{w}\|_2 = 1$. We have $\boldsymbol{w}^\top \Pi_G \boldsymbol{\xi} = \sum_{i\in[N]} (\Pi_G \boldsymbol{w})_i \xi_i$ where the terms of the last sum are independent. The Bernstein inequality yields

$$\mathbb{P}\big(\boldsymbol{w}^\top \Pi_G \boldsymbol{\xi} \leqslant \sigma_\xi \sqrt{2t} + b_\xi t \|\Pi_G \boldsymbol{w}\|_\infty\big) \geqslant 1 - e^{-t}, \qquad t \geqslant 0. \tag{2.38}$$

Let V_G be the image of the unit ball of $\mathbb{R}^N$ by Π_G and let $\mathcal{N}_G = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_M\}$ be an ε-net of V_G for $\varepsilon = 1/4$. The set V_G being included in the unit ball of a linear space of dimension k, [vH16, Lemma 5.13] shows that $M \leqslant 12^k$. Define

$$\boldsymbol{u}_m = \operatorname*{arg\,min}_{u \in V_G} \{\|\boldsymbol{u}\|_\infty : \|\boldsymbol{u} - \boldsymbol{w}_m\|_2 \leq 1/4\}, \qquad m = 1, \dots, M.$$

Using (2.38) and the union bound, we get

$$\mathbb{P}\Big(\forall m \in [M],\ \boldsymbol{u}_m^\top \Pi_G \boldsymbol{\xi} \leqslant \sigma_\xi \sqrt{2(t + \log M)} + b_\xi(t + \log M)\|\boldsymbol{u}_m\|_\infty\Big) \geqslant 1 - e^{-t}, \qquad t \geqslant 0,$$

where we used the fact that $\Pi_G \boldsymbol{u}_m = \boldsymbol{u}_m$. Let $\boldsymbol{w}$ be an arbitrary vector from V_G. Let $\boldsymbol{w}_m$ be any point from the net $\mathcal{N}_G$ such that $\|\boldsymbol{w} - \boldsymbol{w}_m\|_2 \leq 1/4$. We have

$$\|\boldsymbol{u}_m\|_\infty \leq \|\boldsymbol{w}\|_\infty \qquad \text{and} \qquad \|\boldsymbol{w} - \boldsymbol{u}_m\|_2 \leq 1/2.$$

This implies that, with probability at least $1 - e^{-t}$, for any vector $\boldsymbol{w} \in V_G$,

$$\begin{aligned}
\boldsymbol{w}^\top \Pi_G \boldsymbol{\xi} &= (\boldsymbol{w} - \boldsymbol{u}_m)^\top \Pi_G \boldsymbol{\xi} + \boldsymbol{u}_m^\top \Pi_G \boldsymbol{\xi} \\
&\leqslant \|\boldsymbol{w} - \boldsymbol{u}_m\|_2 \|\Pi_G \boldsymbol{\xi}\|_2 + \boldsymbol{u}_m^\top \Pi_G \boldsymbol{\xi} \\
&\leqslant (1/2)\|\Pi_G \boldsymbol{\xi}\|_2 + \sigma_\xi \sqrt{2(t + \log M)} + b_\xi(t + \log M)\|\boldsymbol{u}_m\|_\infty \\
&\leqslant (1/2)\|\Pi_G \boldsymbol{\xi}\|_2 + \sigma_\xi \sqrt{2(t + \log M)} + b_\xi(t + \log M)\|\boldsymbol{w}\|_\infty.
\end{aligned}$$

Since this inequality is valid for any vector $\boldsymbol{w}$ in the image of the unit ball ball by Π_G, it is also valid for $\boldsymbol{w} = \boldsymbol{v}/\|\Pi_G \boldsymbol{v}\|_2$. Replacing this into the last display, then multiplying the two sides of the inequality by $\|\Pi_G \boldsymbol{v}\|_2$, we get the first claim of the lemma.

For the second claim, we take $v = \boldsymbol{\xi}$ to get

$$\|\Pi_G \boldsymbol{\xi}\|_2^2 \leqslant 2\sigma_\xi \sqrt{2(t + \log M)}\|\Pi_G \boldsymbol{\xi}\|_2 + 2b_\xi(t + \log M)\|\Pi_G \boldsymbol{\xi}\|_\infty$$

$$\leqslant (1/2)\|\Pi_G \boldsymbol{\xi}\|_2^2 + 4\sigma_\xi^2(t+\log M) + 2b_\xi(t+\log M)\|\Pi_G\boldsymbol{\xi}\|_\infty,$$

where we used the inequality $2uv \leqslant (1/2)u^2 + 2v^2$ for all $u, v \in \mathbb{R}$. Rearranging the terms of the last display, we obtain the claim of the lemma. □

Lemma 15. *Let $\xi_1, \dots, \xi_N$ be independent zero-mean random variables satisfying the (σ_ξ^2, b_ξ)-Bernstein condition with $\sigma_\xi^2, b_\xi > 0$. Let $\{\mathcal{A}_j : j \in J\}$ be families of subsets of $\{1, \dots, N\}$ such that $N_j = \min_{A\in\mathcal{A}_j} |A| \geqslant N_0$. With probability at least $1 - e^{-t}$, we have*

$$\max_{A\in\mathcal{A}_j} \frac{1}{|A|}\Big|\sum_{\ell\in A}\xi_\ell\Big| \leqslant \sigma_\xi\sqrt{\frac{2(t+\log(2|J|))}{N_0} + \frac{2\log|\mathcal{A}_j|}{N_j}} + \frac{b_\xi(t+\log(2|J|))}{N_0} + \frac{b_\xi\log|\mathcal{A}_j|}{N_j}\ ;\ \forall j \in [J].$$

Proof. Using the version of Lemma 12 of the Bernstein inequality, we find

$$\mathbb{P}\Big(\Big|\sum_{\ell\in A}\xi_\ell\Big| \leqslant \sigma_\xi\sqrt{2|A|t} + b_\xi t\Big) \geqslant 1 - 2e^{-t},$$

which yields

$$\mathbb{P}\Big(\frac{1}{|A|}\Big|\sum_{\ell\in A}\xi_\ell\Big| \leqslant \sigma_\xi\sqrt{\frac{2t}{|A|}} + \frac{b_\xi t}{|A|}\Big) \geqslant 1 - 2e^{-t}.$$

It follows from the last display and the union bound that

$$\mathbb{P}\Big(\max_{A\in\mathcal{A}_j}\frac{1}{|A|}\Big|\sum_{\ell\in A}\xi_\ell\Big| \leqslant \sigma_\xi\sqrt{\frac{2(t+\log|\mathcal{A}_j|)}{N_j}} + \frac{b_\xi(t+\log|\mathcal{A}_j|)}{N_j}\Big) \geqslant 1 - 2e^{-t}.$$

Taking the union bound over $j \in J$, we get that with probability at least $1 - e^{-t}$,

$$\max_{A\in\mathcal{A}_j}\frac{1}{|A|}\Big|\sum_{\ell\in A}\xi_\ell\Big| \leqslant \sigma_\xi\sqrt{\frac{2(t+\log(2|J|)+\log|\mathcal{A}_j|)}{N_j}} + \frac{b_\xi(t+\log(2|J|)+\log|\mathcal{A}_j|)}{N_j}; \forall j \in [J].$$

This completes the proof of the lemma. □

Chapter 3

Graphon estimation in bipartite graphs under relaxed independence assumption

3.1 Introduction

Let us consider a bipartite graph with labeled edges. This means that the vertices of the graph are split into two parts, of cardinalities n and m, respectively, so that there is no edge between two vertices belonging to the same part. Thus, only vertices lying in different parts may be connected by an edge. Such a graph is naturally encoded by its adjacency matrix, which is an $n \times m$ matrix henceforth denoted by $\mathbf{H}$ such that $H_{ij} = 1$ if the the ith vertex of the first part is connected to the jth vertex of the second part, otherwise $H_{i,j} = 0$. In a more general setting that will be explained below, one can assume that the matrix $\mathbf{H}$ has real entries from $[0, 1]$ representing the "labels" of the edges of the graph.

To give a concrete example, let the vertices of the first part of the graph be workers and those of the second part be firms. A worker is connected to a firm if they are employed by the firm. Clearly, this graph evolves over time, but we consider here its state at a given time instant. It could be tempting to apply the methodology of the previous chapter in this setting, but it can be easily seen that this requires relaxing some of the assumptions considered therein. Indeed, at a given time instant, a worker is employed by at most one firm, which means that every row of $\mathbf{H}$ contains only one non-zero entry. This implies that no assumption of independence, even conditional, can not be envisioned for modeling this setting. Having this and other similar examples in mind, we introduce an alternative version of the assumption that the graph is generated by a graphon.

Assumption 4. We consider a function $W^* : [0, 1]^2 \to [0, +\infty[$ called the graphon and two random vectors $\boldsymbol{U} = (U_1, \dots, U_n)$ and $\boldsymbol{V} = (V_1, \dots, V_m)$ that satisfy

1. $U_1, \dots, U_n, V_1, \dots, V_m$ are independent and drawn from the uniform distribution on $[0, 1]$.
2. Conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the rows of the matrix $\mathbf{H}$ are independent.
3. Each row of $\mathbf{H}$ sum to one and $\mathbb{E}[H_{i,j}|\boldsymbol{U}, \boldsymbol{V}] = \dfrac{W^*(U_i, V_j)}{\sum_{\ell=1}^{m} W^*(U_i, V_\ell)}$.

In this assumption, as mentioned in the previous chapter, U_i's and V_j's are unobserved features of the vertices. In the example of workers and firms, one can think of U_i as the unobserved efficiency of the i-th worker, whereas V_j is the attractiveness of the jth firm, which is not directly observable. Part 2 of Assumption 4, referred to as the *partial independence* assumption, indicates that if we were to observe the values of the latent vectors $\boldsymbol{U}$ and $\boldsymbol{V}$, then each worker would choose the firm independently of the other workers. However, there is no assumption on the (conditional) independence of the choices made by a worker between different firms. This allows us to cover the "many to one" situation described above, which means that many workers may be employed by one firm but only one firm can be the employer of a worker.

The focus of this chapter is on estimating the conditional mean, given the latent variables $\boldsymbol{U}$ and $\boldsymbol{V}$, of the random matrix $\mathbf{H}$ as well as on estimating the graphon W^*.

Besides the worker-firm example described above, the model under consideration can be used in the following situations.

1. The entries of $\mathbf{H}$ live in $\{0,1\}$. Condition 3 of Assumption 4 is satisfied if each row of $\mathbf{H}$ has one entry equal to one, the others vanish. This corresponds to the matching setting, where a worker, for instance, has to choose one and only one firm. In this case, the matrix $\mathbf{H}$ is sparse and has exactly n nonzero entries. This is equivalent to assume that conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the rows of $\mathbf{H}$ are independent and drawn from the multinomial distribution with parameters $(1, m, \boldsymbol{\Theta}_i^*)$ where $\boldsymbol{\Theta}_i^*$ lies in the probability simplex and its j coordinate is proportional to $W^*(U_i, V_j)$.

2. Consider that the setting of the previous paragraph has been repeated N times, resulting in the observation of N adjacency matrices $\mathbf{H}_1, \dots, \mathbf{H}_N$. We can set $\mathbf{H} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{H}_k$. This may model the situation where we observe the employment graph at N different time instances.

3. Let us assume that shareholders invest in different investment classes. Each shareholder has a budget allocated to invest, and we observe the fraction of this budget invested in each class. Here, entries of $\mathbf{H}$ are in $[0,1]$, they measure the fraction of the budget invested by each shareholder in each class. One row of $\mathbf{H}$ corresponds to one shareholder. A typical example is when conditionally to $(\boldsymbol{U}, \boldsymbol{V})$, the ith row of $\mathbf{H}$ is drawn from the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha}_i)$ with parameter $\boldsymbol{\alpha}_i = \big(W^*(U_i, V_1), \dots, W^*(U_i, V_m)\big)$.

Our goal is twofold. First, we aim to estimate the mean matrix of $\mathbf{H}$ and provide a risk bound for our estimation method. In this part, we do not consider part 3 of Assumption 4, but instead replace it with the following assumption: the sum of each row $\boldsymbol{H}_i$ of $\mathbf{H}$ is smaller than a positive parameter $\rho_{\mathbf{H}}$. Our objective is to investigate the behavior of the risk of our estimator as a function of the model parameters, namely n and m, as well as $\rho_{\mathbf{H}}$ and the noise level denoted by $\rho_{\boldsymbol{\Sigma}}$, which satisfies

$$\|\boldsymbol{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\boldsymbol{\Sigma}} \quad \forall i = 1 \dots n \tag{3.1}$$

where $\boldsymbol{\Sigma}_i = \mathbb{E}[\boldsymbol{H}_i \boldsymbol{H}_i^\top] - \mathbb{E}[\boldsymbol{H}_i]\mathbb{E}[\boldsymbol{H}_i]^\top$ is the covariance matrix of the ith row of $\mathbf{H}$. We will often assume that the matrix $\boldsymbol{\Theta}^* = \mathbb{E}[\mathbf{H}]$ has also rows whose sum is less than $\rho_{\boldsymbol{\Theta}} > 0$. This parameter may also appear in the upper bounds of the risk. Notice that $\rho_{\boldsymbol{\Theta}}$ could be much smaller than $\rho_{\mathbf{H}}$. For example, if $\boldsymbol{H}_i$ is multinomial with parameters $(1, m, (1/m, \dots, 1/m))$, then $\rho_{\mathbf{H}} = 1$ and $\rho_{\boldsymbol{\Theta}} = 1/m$. The second part of this work focuses on estimating the graphon W^* when $\mathbf{H}$ fully satisfies Assumption 4 as well as (3.1). Moreover we only consider the class of piecewise constant graphon W^*. Notice that part 3 of Assumption 4 poses identifiability problems. Indeed, conditionally on $(\boldsymbol{U}, \boldsymbol{V})$, W^* and CW^*, where C is a positive constant, will generate the same matrix $\mathbf{H}$. In addition, permuting the coordinates V_j and U_i will also

generate the same matrix $\mathbf{H}$. These identifiability problems are treated in Subsection 3.3.1, and they are important to understand the way of measuring the risk of our estimator.

Estimation procedure In the initial stage of our work, our objective is to estimate the mean matrix $\mathbf{\Theta}^* = \mathbb{E}[\mathbf{H}]$. To accomplish this, we employ a least squares approach using constant-by-block matrices, as proposed in Chapter 2. The number of blocks is assumed to be known and serves as a hyperparameter in our model, influencing the upper bounds. While adaptive methods exist for handling unknown block numbers, such as aggregation by exponential weights discussed in Subsection 2.3.2, we do not delve into them in this study. In the subsequent stage, we shift our focus to estimating the piecewise constant graphon W^* and leverage the estimator of the mean matrix to derive an estimator for W^*.

Our contribution Our work makes a significant contribution by providing a sample risk bound for the least squares method employed in estimating the mean of a random matrix under the relaxed independence assumption, where the entries of the observed matrix are not necessarily independent, but the rows are assumed to be. To tackle the least squares problem, we use constant-by-block constraints, assuming that the number of blocks is known. Additionally, we establish an upper bound for estimating a bivariate graphon function under piecewise constant regularity within the framework of Assumption 4, which incorporates unobserved latent variables. This framework allows for modeling a wide range of random matrices, particularly adjacency matrices of bipartite graphs, with constraints on edge labels.

Notation For two $n \times m$ matrices $\mathbf{B}$ and $\mathbf{D}$, the inner product is defined as

$$\langle \mathbf{B}, \mathbf{D} \rangle = \mathrm{tr}(\mathbf{B}\mathbf{D}^\top) = \sum_{i=1}^{n} \sum_{j=1}^{m} B_{ij} D_{ij},$$

and we denote by $\|\mathbf{B}\|_{\mathsf{F}} = \sqrt{\langle \mathbf{B}, \mathbf{B} \rangle}$ the Frobenius norm of the matrix $\mathbf{B}$. We denote $\|\mathbf{B}\|_{p,\infty} = \max_{i \in [n]} \|\boldsymbol{B}_i\|_p$ where $p \in \{1, 2\}$. For any integer $N \geqslant 1$, the notation $[N]$ will refer to the set $\{1, \ldots, N\}$.

3.2 Estimators of the mean matrix and finite sample risk bound

In this section, we introduce the least square estimator for the mean matrix $\mathbf{\Theta}^*$, which is computed from the matrix $\mathbf{H}$ where the rows of $\mathbf{H}$ consist of independent observed vectors $\boldsymbol{H}_i^\top$. Furthermore, we provide an upper-bound on the risk associated with the least square estimator.

3.2.1 Definition of the least square estimator

The definition of the least square estimator is essentially identical to the one provided in Equation (2.2). However, let's revisit the framework for the sake of clarity. The approach involves approximating the observed matrix $\mathbf{H}$ using a constant-by-block matrix.

For positive integers n_0, n, K satisfying $Kn_0 \leqslant n$, and $K \geqslant 2$, we define the set $\mathcal{Z}(n, K, n_0)$ as follows:

$$\mathcal{Z}(n, K, n_0) = \left\{ \mathbf{Z} \in \{0,1\}^{n \times K} : \mathbf{Z}\mathbf{1}_K = \mathbf{1}_n \text{ and } \min_{k \in [K]} \mathbf{1}_n^\top \boldsymbol{Z}_k \geqslant n_0 \right\}.$$

The elements of this set can be interpreted as assignment matrices, where each node of the first set with cardinality n is assigned to one (and only one) of the K "communities," subject to the condition that each community has at least n_0 "members." Similarly, we will make use of the set $\mathcal{Z}(m, L, m_0)$ corresponding to assignment matrices for the second set. By convention, we consider the first set to be on the left side, and the second set on the right side. Consequently, elements of $\mathcal{Z}(n, K, n_0)$ will be denoted as $\mathbf{Z}^{\text{user}}$, while elements of $\mathcal{Z}(m, L, m_0)$ will be denoted as $\mathbf{Z}^{\text{item}}$. These matrices represent a biclustering of the bipartite network, where the left clusters are given by the matrix $\mathbf{Z}^{\text{user}}$, and similarly, $\mathbf{Z}^{\text{item}}$ represents the right clusters. Given the observed matrix $\mathbf{H}$, the least square estimator is defined by

$$(\widehat{\mathbf{Q}}, \widehat{\mathbf{Z}}^{\text{user}}, \widehat{\mathbf{Z}}^{\text{item}}) \in \arg \min_{\substack{\mathbf{Q} \in [0,1]^{K \times L} \\ \mathbf{Z}^{\text{user}} \in \mathcal{Z}(n,K,n_0) \\ \mathbf{Z}^{\text{item}} \in \mathcal{Z}(m,L,m_0)}} \left\| \mathbf{H} - \mathbf{Z}^{\text{user}} \mathbf{Q} (\mathbf{Z}^{\text{item}})^\top \right\|_{\mathsf{F}}^2. \tag{3.2}$$

Here, $\mathbf{Z}^{\text{user}} \mathbf{Q} (\mathbf{Z}^{\text{item}})^\top$ is a $n \times m$ constant-by-block matrix. The idea is thus to find the constant-by-block matrix that is the closest to $\mathbf{H}$ in the metric induced by the Frobenius norm, where the blocs are given by the matrices $\mathbf{Z}^{\text{user}}$ and $\mathbf{Z}^{\text{item}}$, and the sizes of blocks are at least n_0 and m_0.

These estimators computed by (3.2) lead to the constant-by-block least squares estimator of $\boldsymbol{\Theta}^*$ defined by $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{Z}}^{\text{user}} \widehat{\mathbf{Q}} (\widehat{\mathbf{Z}}^{\text{item}})^\top$. One can write $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}$ in the following alternative way. Let us consider the class of constant-by-block matrices

$$\mathcal{T} = \left\{ \boldsymbol{\Theta} = \mathbf{Z}^{\text{user}} \mathbf{Q} (\mathbf{Z}^{\text{item}})^\top \in [0,1]^{n \times m} : \ (\mathbf{Q}, \mathbf{Z}^{\text{user}}, \mathbf{Z}^{\text{item}}) \in [0,1]^{K \times L} \times \mathcal{Z}_{n,K,n_0} \times \mathcal{Z}_{m,L,m_0} \right\}.$$

The estimator $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}$ is a solution to

$$\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} \in \arg \min_{\boldsymbol{\Theta} \in \mathcal{T}} \|\mathbf{H} - \boldsymbol{\Theta}\|_{\mathsf{F}}^2. \tag{3.3}$$

We aim at bounding the error of the estimator $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}$.

3.2.2 Risk bound of the least square estimator

We present an upper bound on the error of the least squares estimator, which mirrors the one obtained in Theorem 6. This bound illustrates the behavior of the error as n and m grow large, while keeping $\rho_{\mathbf{\Theta}}, \rho_{\mathbf{H}}$, and $\rho_{\mathbf{\Sigma}}$ of the order of one.

Proposition 5. *Let n, m, n_0, m_0, L, K be positive integers such that $L \geqslant 2$, $K \geqslant 2$, $1 \leqslant n_0 \leqslant n$ and $1 \leqslant m_0 \leqslant m$. Let $\mathbf{H} \in [0,1]^{n \times m}$ be an $n \times m$ random matrix with independent rows such that $\|\mathbf{H}\|_{1,\infty} \leqslant \rho_{\mathbf{H}}$. Assume that the mean matrix $\mathbf{\Theta}^*$ satisfies $\|\mathbf{\Theta}^*\|_{1,\infty} \leqslant \rho_{\mathbf{\Theta}}$ and the covariance matrix $\mathbf{\Sigma}_i^*$ of $\boldsymbol{H}_i$ is such that $\|\mathbf{\Sigma}_i^*\|_{\mathrm{op}} \leq \rho_{\mathbf{\Sigma}}$ for every $i \in [n]$ and for some $\rho_{\mathbf{\Sigma}} \geqslant 0$. The least-squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ defined by (3.3) satisfies the exact oracle inequality*

$$\frac{\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \inf_{\mathbf{\Theta} \in \mathcal{T}} \frac{\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + \big(24\rho_{\mathbf{\Sigma}} + 6\rho_{\mathbf{\Theta}}\big)^{1/2} r_{n,m}(K,L),$$

provided that $\psi_{n,m}(n_0, m_0) = \frac{\log(ne/n_0)}{m_0} + \frac{\log(me/m_0)}{n_0} \leqslant \frac{\rho_{\mathbf{\Sigma}}}{8(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})^2}$ and the remainder factor is

$$r_{n,m}(K,L) = \Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2}.$$

This upper bound is particularly meaningful when the parameters $\rho_{\mathbf{\Sigma}}$, $\rho_{\mathbf{\Theta}}$, and $\rho_{\mathbf{H}}$ are bounded when n and m go to infinity. If, for instance, $\rho_{\mathbf{\Theta}}$ is proportional to m, then the upper bound will not converge to 0 as n and m increase. For next section, in view of the previous remark, we will focus on the case where $\rho_{\mathbf{\Theta}} = \rho_{\mathbf{H}} = 1$, this assumption being true in the frameworks we are interested in. It is worth noting that the assumption $\|\mathbf{H}\|_{1,\infty} \leqslant \rho_{\mathbf{H}}$ ensures that $\|\mathbf{\Theta}^*\|_{1,\infty} \leqslant \rho_{\mathbf{H}}$ as well. However, it should be mentioned that in some cases, $\|\mathbf{\Theta}^*\|_{1,\infty}$ may be smaller than $\|\mathbf{H}\|_{1,\infty}$. If we further assume in Proposition 5 that $\|\mathbf{\Theta}^*\|_{\infty} \leqslant \rho_{\infty}$, we can easily adapt the proof to obtain the following upper bound

$$\inf_{\mathbf{\Theta} \in \mathcal{T}} \frac{\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + \Big(24\rho_{\mathbf{\Sigma}} + 6\rho_{\infty}(\rho_{\mathbf{\Theta}} + \rho_{\mathbf{H}})\Big)^{1/2} r_{n,m}(K,L).$$

This bound can be compared to the bound obtained in the independent edges framework, as given in Theorem 6, which applies to bounded labels (see Table 3.1).

The next corollary is a straightforward consequence of Proposition 5. It corresponds to the particular setting where the rows of $\mathbf{H}$ sum to one. In this case, we have the following inequalities

$$\rho_{\mathbf{\Sigma}} \leqslant \rho_{\infty} \leqslant \rho_{\mathbf{\Theta}} = \rho_{\mathbf{H}} = 1$$

This includes the case described in the introduction where each row $\boldsymbol{H}_i$ has only one entry equal to 1, and the others are null. This case, known as the multinomial model, models a matching scenario where individuals from the left-hand side have to choose one and only one

Assumption	Full independence	Partial Independence
Bounded data	b	$\rho_{\mathbf{E}} = \rho_{\mathbf{\Theta}} + \rho_{\mathbf{H}}$
Bounded expectation	ρ	ρ_∞
Bounded variance	σ^2	$\rho_{\mathbf{\Sigma}}$
Upper bound factor provided that	$\big(\sigma^2 + b\rho\big)^{1/2}$ $\psi_{n,m}(n_0, m_0)\big) \lesssim (\sigma/b)^2$	$\big(\rho_{\mathbf{\Sigma}} + \rho_{\mathbf{E}}\rho_\infty\big)^{1/2}$ $\psi_{n,m}(n_0, m_0) \lesssim \rho_{\mathbf{\Sigma}}/\rho_{\mathbf{E}}^2$

Table 3.1: Comparison of the parameters and the factors appearing in the upper bound for full independence and partial independence assumptions.

item from the right-hand side.

Corollary 3. Let n, m, n_0, m_0, L, K be positive integers such that $L \geqslant 2$, $K \geqslant 2$, $1 \leqslant n_0 \leqslant n$ and $1 \leqslant m_0 \leqslant m$. Let $\mathbf{H} \in [0,1]^{n \times m}$ be an $n \times m$ random matrix with independent rows such that each row sum to one and has a covariance matrix $\mathbf{\Sigma}_i$ satisfying $\|\mathbf{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\mathbf{\Sigma}}$. We also assume that $\|\mathbf{\Theta}^*\|_\infty \leqslant \rho_\infty$. The least-squares estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ defined by (3.3) satisfies the exact oracle inequality

$$\frac{\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2}}{\sqrt{nm}} \leqslant \inf_{\mathbf{\Theta} \in \mathcal{T}} \frac{\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + \big(48\rho_{\mathbf{\Sigma}} + 6\rho_\infty\big)^{1/2} r_{n,m}(K, L),$$

provided that $\psi_{n,m}(n_0, m_0) \leqslant \rho_{\mathbf{\Sigma}}$, with the remainder factor

$$r_{n,m}(K, L) = \Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2}.$$

Notice that $\|\mathbf{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_\infty$, which can simplifies the upper bound obtained in Corollary 3, the factor before $r_{n,m}(K, L)$ becomes $8\sqrt{\rho_\infty}$, and corresponds to the optimal upper bound obtained in the multinomial model in terms of the noise and the sparsity parameters. For the reweighed multinomial model (second example presented in the introduction), the upper-bound is sub-optimal in term of the variance noise, because $\rho_{\mathbf{\Sigma}}$ is of the order of $1/N$ while ρ_∞ behaves as a constant.

3.3 Estimators of the graphon and risk bound

Having obtained the estimation of the mean matrix $\mathbf{\Theta}^*$, our next objective is to construct an estimator for W^* within the framework of Assumption 4. Additionally, we will address the identifiability issues mentioned in the introduction and provide an upper bound for the risk of this graphon estimator.

3.3.1 Identifiability and evaluation of the estimation

In this subsection, we define the estimator of W^* when the matrix $\mathbf{H}$ is generated according to some re-scaled graphon W^*. Assume that $\mathbf{H}$ has independent rows $\boldsymbol{H}_i$ that sum to one, so $\sum_{j=1}^m H_{ij} = 1$ and let consider the model

$$\mathbb{E}[\mathbf{H}|\boldsymbol{U}, \boldsymbol{V}] = \boldsymbol{\Theta}^* \quad \text{with} \quad \Theta^*_{ij} = \frac{W^*(U_i, V_j)}{\sum_{\ell=1}^m W^*(U_i, V_\ell)} \tag{3.4}$$

where $W^* : [0,1] \to \mathbb{R}_+^*$ is a graphon, U_i and V_j are iid sequences of uniform variables on $[0,1]$. In the context of bipartite graphs, the matrix $\mathbf{H}$ represents an adjacency matrix. We associate an unobserved latent variable U_i with each vertex on the left-hand side, and similarly, we assign the variables V_j to the vertices on the right-hand side. They represent unobservable characteristics, and they are assumed to be independent.

The goal is to construct an estimator of W^* based on the observation of the matrix $\mathbf{H}$. However, it is important to note that model (3.4) is invariant under the rescaling of W^* by a positive constant. Moreover, if there exists τ, $\tau_1, \dots, \tau_m$ some bijections from $[0,1]$ to $[0,1]$ that preserve the Lebesgue measure and such that

$$\frac{W(x, y_1)}{\sum_{\ell=1}^m W(x, y_\ell)} = \frac{W'(\tau(x), \tau_1(y_1))}{\sum_{\ell=1}^m W'(\tau(x), \tau_\ell(y_\ell))} \qquad \text{for all } x, y_1 \dots, y_m \in [0,1] \tag{3.5}$$

then, matrices $\mathbf{H}$ and $\mathbf{H}'$ respectively generated from W and W' according to model (3.4) have the same distribution. That is, the only observation of $\mathbf{H}$ will not allow to discriminate between W and W'. It is important to well understand this claim to avoid identifiability issues. One can rewrite (3.5) in a simpler way.

Proposition 6. *If there exists τ, $\tau_1, \dots, \tau_m$ some bijections from $[0,1]$ to $[0,1]$ that preserve the Lebesgue measure and such that for all $x, y_1 \dots, y_m \in [0,1]$, equation (3.5) is satisfied, then*

$$W(x, y) = \nu(x) W'(\pi_1(x), \pi_2(y)) \quad \textit{for all} \quad x, y \in [0,1]. \tag{3.6}$$

where ν is a positive function, and π_1, π_2 are some measure preserving bijections of $[0,1]$.

Notice that the reciprocal is also true, the proof is straightforward. At the sight of Proposition 6, we say that two graphons W and W' are weakly isomorphic if (3.6) is satisfied. Then, as it is impossible to discriminate between two graphons that are equal up to a multiplicative function that depends only on x, we will define the class of graphons

$$\mathscr{C} = \left\{ W, I_W(x) = 1/m \ \ \forall x \in [0,1] \right\}$$

where $I_W(x) = \int_0^1 W(x, y) dy$. Assume that $W^* \in \mathscr{C}$, otherwise we normalize it, and it will not

change the model (3.4). This normalization is motivated by the the fact that if $W^* \in \mathscr{C}$, then

$$\mathbb{E}\Big[\sum_{\ell=1}^{m} W^*(U_i, V_\ell)|U_i\Big] = mI_{W^*}(U_i) = 1,$$

and, intuitively, we should have $\mathbb{E}[\mathbf{H}|U_i, V_j] \approx W^*(U_i, V_j)$. Now, the good distance on the class $\mathscr{C}$ we choose for measuring the quality of an estimator of W^* is

$$\begin{aligned}\delta(W', W) &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \Big(\iint_{[0,1]^2} \big|W'(\tau_1(u), \tau_2(v)) - W(u, v)\big|^2 du\, dv\Big)^{1/2} \\ &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \big\|W' \circ (\tau_1 \otimes \tau_2) - W\big\|_{\mathbb{L}^2}\end{aligned}$$

or equivalently, the distance between two graphons that are not necessarily in $\mathscr{C}$ is define as

$$\begin{aligned}\delta(W', W) &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \Big(\iint_{[0,1]^2} \Big|\frac{W'(\tau_1(u), \tau_2(v))}{mI_{W'}(\tau_1(u))} - \frac{W(u, v)}{mI_W(u)}\Big|^2 du\, dv\Big)^{1/2} \\ &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \Big\|\frac{W' \circ (\tau_1 \otimes \tau_2)}{mI_{W'} \circ \tau_1} - \frac{W}{mI_W}\Big\|_{\mathbb{L}^2}\end{aligned}$$

The distance δ has important properties with regard to identifiability. Specifically, if two graphons are weakly isomorphic, which means that there exists a positive function ν, and measure-preserving bijections τ_1 and τ_2 such that for almost all $x, y \in [0, 1]$, we have $W'(x, y) = \nu(x)W(\tau_1(x), \tau_2(y))$, then $\delta(W, W') = 0$.

Having established the distance metric, we now proceed to define the graphon estimator. Similar to the previous chapter, we define the function $W_{\mathbf{\Theta}}(x, y)$ for any matrix $\mathbf{\Theta} \in \mathbb{R}^{n\times m}$ as $W_{\mathbf{\Theta}}(x, y) = \Theta_{\lceil nx\rceil,\lceil my\rceil}$. The least squares estimator, denoted by $\widehat{W}^{\mathsf{LS}}$, is then define by $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}$. Notice that

$$I_{W_{\mathbf{\Theta}}}(x) = \int_0^1 W_{\mathbf{\Theta}}(x, y)dy = \frac{1}{m}\sum_{j=1}^{m} \Theta_{\lceil nx\rceil,j}.$$

Then, $W_{\mathbf{\Theta}} \in \mathscr{C}$ if and only if $\mathbf{\Theta}$ is left stochastic, that is $\sum_{j=1}^{m} \Theta_{ij} = 1$ for all $i \in [n]$ (all its row sum to one). Hence, $W_{\mathbf{\Theta}^*} \in \mathscr{C}$ because of model (3.4).The next lemma guarantees that if $\mathbf{H}$ has rows that sum to one, then $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ will also have rows that sum to one, and $\widehat{W}^{\mathsf{LS}} \in \mathscr{C}$.

Lemma 16. *If $\mathbf{H} \in (\mathbb{R}_+)^{n\times m}$ is left-stochastic, meaning that all its rows sum to one, then the least square estimator $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ define by (3.3) is also left-stochastic.*

For an estimated graphon $W_{\widehat{\mathbf{\Theta}}}$ defined above, such that $\widehat{\mathbf{\Theta}}$ is left-stochastic (that is $W_{\widehat{\mathbf{\Theta}}} \in \mathscr{C}$), the error of estimation can be easily related to the error, measured by the Frobenius norm, of estimation of the matrix $\mathbf{\Theta}^*$. Indeed, one easily checks that $\|W_{\widehat{\mathbf{\Theta}}} - W_{\mathbf{\Theta}^*}\|_{\mathbb{L}_2} = \frac{1}{\sqrt{nm}}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}$,

which leads to

$$\delta(W_{\widehat{\mathbf{\Theta}}}, W^*) \leqslant \|W_{\widehat{\mathbf{\Theta}}} - W_{\mathbf{\Theta}^*}\|_{\mathbb{L}_2} + \delta(W_{\mathbf{\Theta}^*}, W^*) \leqslant \frac{\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}}{\sqrt{nm}} + \delta(W_{\mathbf{\Theta}^*}, W^*). \tag{3.7}$$

The equation (3.7) provides a decomposition that divides the error of graphon estimation into two distinct components: the error of estimating the conditional mean matrix $\mathbf{\Theta}^*$, and the bias of approximating W^* with a piecewise constant function $W_{\mathbf{\Theta}^*}$. The former component is influenced by the estimation method used and the probabilistic assumptions made about the noise, and has been examined in the previous section. Meanwhile, the latter component is dependent solely on the "smoothness properties" of the graphon. A forthcoming result will enable us to assess this component.

3.3.2 Risk bound for piecewise constant graphons

In this subsection, our focus is on deriving an upper bound for the risk of the least square estimator of the graphon W^*, given the assumption that W^* is piecewise constant.

Proposition 7. *For $i \in [n]$ and $j \in [m]$, let Θ^*_{ij} define as follow*

$$\Theta^*_{ij} = \frac{W^*(U_i, V_j)}{\sum_{\ell=1}^m W^*(U_i, V_\ell)}$$

where $W^* : [0,1]^2 \to [A, B]$ for some A, B such that $0 \leqslant A < B \leqslant +\infty$. Assume firstly that $W^* \in \mathscr{C}$, that is $I_{W^*}(x) := \int_0^1 W^*(x, y)dy = 1/m$ and secondly that for $0 = a_0 < \ldots < a_K = 1$ and $0 = b_0 < \ldots < b_L = 1$, the function W^* is constant on each rectangle $[a_k, a_{k+1}[\times[b_\ell, b_{\ell+1}[$. If we define $W_{\mathbf{\Theta}^*} : [0,1]^2 \to [A, B]$ by $W_{\mathbf{\Theta}^*}(u, v) = \Theta^*_{i,j}$ for all $(u, v) \in [(i-1)/n, i/n[\times[(j-1)/n, j/n[$, then

$$\mathbb{E}[\delta(W_{\mathbf{\Theta}^*}, W^*)] \leqslant \frac{(B-A)}{\sqrt{2}}\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2} + B\sqrt{\frac{3}{m}}$$

provided that $\bar{w} \leqslant \frac{1}{4}e^{0.045m}$ where $\bar{w} = \sum_{\ell=1}^L \frac{1}{w^{(2)}_\ell}$ and $w^{(2)}_\ell = b_\ell - b_{\ell-1}$.

The Bound obtained is quite similar to the one given in Proposition 2. To be more accurate, the only difference lies in the presence of the second term $\sqrt{\frac{B^2}{m}}$. It appears because of the normalisation given by the graphon's class $\mathscr{C}$. But one have that this term is much smaller than the first one when A and B are fixed, then the obtained upper bounds are essentially the same for full independence and partial Independence setting. Note that there is an additional assumption here, we enforce $\bar{w}$ to be not too large, that is, the size of the intervals $[b_\ell, b_{\ell+1}[$ is not too small. For example if $m = 315$, then this condition is satisfied as long as $\min_{\ell \in [L]} |b_\ell - b_{\ell-1}| \geqslant 10^{-3}$. Now we have all the tools to upper bound the risk estimation of $\widehat{W}^{\mathsf{LS}}$.

Theorem 10. *Let $\mathbf{H} \in [0,1]^{n\times m}$ be a $n \times m$ random matrix satisfying*

$$\mathbb{E}[\mathbf{H}|\boldsymbol{U},\boldsymbol{V}] = \boldsymbol{\Theta}^* \quad \text{with} \quad \Theta^*_{ij} = \frac{W^*(U_i,V_j)}{\sum_{\ell=1}^m W^*(U_i,V_\ell)}$$

with some graphon $W^* : [0,1]^2 \to [0,\rho] \in \mathscr{C}$, that is $I_{W^*}(x) = \int_0^1 W^*(x,y)dy = 1/m$. Assume that each row of $\mathbf{H}$ sum to one, its covariance matrix $\boldsymbol{\Sigma}_i$ satisfies $\|\boldsymbol{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\boldsymbol{\Sigma}} \leqslant 1$ and its conditional mean matrix $\boldsymbol{\Theta}^*$ satisfies $\|\boldsymbol{\Theta}^*\|_\infty \leqslant \rho$. Assume that the graphon W^* is (K,L)-piecewise constant, meaning that for some integers $K, L \geqslant 2$ and for $0 = a_0 < \ldots < a_K = 1$, $0 = b_1 < \ldots < b_L = 1$, such that

$$\Delta^{(K)} := \min_{k\in[K]} |a_k - a_{k-1}| \geqslant \frac{8\log(nK)}{n}, \quad \Delta^{(L)} := \min_{\ell\in[L]} |b_\ell - b_{\ell-1}| \geqslant \frac{8\log(mL)}{m},$$

the function W^* is constant on each rectangle $[a_{k-1},a_k[\times[b_{\ell-1},b_\ell[$. Then, the estimator $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}}$ with $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} = \widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K,L]$ defined by (3.3) satisfies

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}},W^*)^2]^{1/2} \leqslant \big(50\rho_{\boldsymbol{\Sigma}} + 6\rho\big)^{1/2}\Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} + 3\rho\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2}$$

provided that $\psi_{n,m}(\Delta^{(K,L)}) = \frac{2\log(2e/\Delta^{(K)})}{m\Delta^{(L)}} + \frac{2\log(2e/\Delta^{(L)})}{n\Delta^{(K)}} \leqslant \rho_{\boldsymbol{\Sigma}}$ and $\bar{w} \leqslant \frac{1}{4}e^{0.045m}$ where $\bar{w} = \sum_{\ell=1}^L \frac{1}{w_\ell^{(2)}}$ and $w_\ell^{(2)} = b_\ell - b_{\ell-1}$.

In order to ease understanding of these results, let give the simple example of the multinomial model, where each row of $\mathbf{H}$ are drawn from a multinomial distribution of parameter $(0,1,\boldsymbol{\Theta}_i)$. Then $\rho_{\boldsymbol{\Sigma}} \leqslant \rho_\infty$ and the upper bound becomes

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}},W^*)^2]^{1/2} \leqslant 8\Big(\frac{3\rho KL}{nm} + \frac{\rho\log K}{m} + \frac{\rho\log L}{n}\Big)^{1/2} + 3\rho\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2}.$$

This upper bound is optimal in term of the noise level and the sparsity parameter ρ. It can be compared to the upper-bound obtain in part 1 of Theorem 8 in the Bernoulli setting. However, this result becomes sub-optimal when applying a re-weighted multinomial model.

3.4 Proofs

3.4.1 Proof of Proposition 5 (risk bound for $\widehat{\Theta}^{\mathsf{LS}}$)

The proof of Proposition 5 presented here is an adaptation of the proof of Theorem 6 in Chapter 2. Let us define

$$\Pi_{\mathcal{T}}(\mathbf{\Theta}^*) = \arg\min_{\mathbf{\Theta}\in\mathcal{T}} \|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}},$$

the best approximation of $\mathbf{\Theta}^*$ in Frobenius norm by a constant-by-block matrix. Note that the matrix $\Pi_{\mathcal{T}}(\mathbf{\Theta}^*)$ has at most KL distinct entries each of which is the average of the entries of a submatrix of $\mathbf{\Theta}^*$. Since $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{Z}}^{\mathrm{user}}\widehat{\mathbf{Q}}(\widehat{\mathbf{Z}}^{\mathrm{item}})^\top$ is the least square estimator, we have

$$\left\|\mathbf{H} - \widehat{\mathbf{\Theta}}^{\mathsf{LS}}\right\|_{\mathsf{F}}^2 \leqslant \left\|\mathbf{H} - \Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\right\|_{\mathsf{F}}^2. \tag{3.8}$$

Let us define the mean-zero "noise" matrix $\mathbf{E} = \mathbf{H} - \mathbb{E}[\mathbf{H}] = \mathbf{H} - \mathbf{\Theta}^*$ and rewrite equation (3.8) in the following form

$$\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \leqslant \|\mathbf{\Theta}^* - \Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\|_{\mathsf{F}}^2 + 2\langle\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*, \mathbf{E}\rangle + 2\langle\mathbf{\Theta}^* - \Pi_{\mathcal{T}}(\mathbf{\Theta}^*), \mathbf{E}\rangle. \tag{3.9}$$

Since the mean of $\mathbf{E}$ is zero, the expectations of the last term in the right-hand side vanishes. To bound the expectation of $\langle\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*, \mathbf{E}\rangle$, we define

$$\widehat{\mathcal{T}} = \left\{\mathbf{\Theta} : \exists \mathbf{Q} \in [0,1]^{K\times L} \text{ such that } \mathbf{\Theta} = \widehat{\mathbf{Z}}^{\mathrm{user}}\mathbf{Q}(\widehat{\mathbf{Z}}^{\mathrm{item}})^\top\right\} \subset \mathcal{T},$$

and let $\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) = \arg\min_{\mathbf{\Theta}\in\widehat{\mathcal{T}}} \|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}}$ be the best Frobenius approximation of $\mathbf{\Theta}^*$ in $\widehat{\mathcal{T}}$. We use the decomposition

$$\langle\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*, \mathbf{E}\rangle = \underbrace{\langle\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*) - \mathbf{\Theta}^*, \mathbf{E}\rangle}_{\Xi_1} + \underbrace{\langle\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*), \mathbf{E}\rangle}_{\Xi_2}. \tag{3.10}$$

Lemma 17. *Under the conditions of Proposition 5, we have*

$$\mathbb{E}(\Xi_1) \leqslant \sqrt{2\rho_{\mathbf{\Sigma}}(n\log K + m\log L + 1)}\,\mathbb{E}[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2} + (4/3)\rho_{\mathbf{\Theta}}(n\log K + m\log L + 1).$$

Proof. The main steps of the proof consist in applying the Bernstein inequality to Ξ_1 for a

fixed $\boldsymbol{\Theta}$ instead of $\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*)$, using the union bound and then integrating the high-probability bound. For the first step, let $\boldsymbol{\Theta} \in [0,1]^{n\times m}$ satisfying $\|\boldsymbol{\Theta}\|_{1,\infty} \leq \rho_{\boldsymbol{\Theta}}$. By definition of the inner product, we have $\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^*, \mathbf{E}\rangle = \sum_{i\in[n]} (\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*)^\top \boldsymbol{E}_i$. The random variables $(\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*)^\top \boldsymbol{E}_i$ are independent and satisfy

$$\begin{aligned}
\big|(\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*)^\top \boldsymbol{E}_i\big| &\leqslant \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{1,\infty} \leqslant 2\rho_{\boldsymbol{\Theta}},\\
\mathbb{E}\big[\big\{(\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*)^\top \boldsymbol{E}_i\big\}^2\big] &= (\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*)^\top \boldsymbol{\Sigma}_i^* (\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*) \leqslant \|\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*\|_2^2 \|\boldsymbol{\Sigma}_i^*\|_{\mathrm{op}} \leqslant \rho_{\boldsymbol{\Sigma}} \|\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*\|_2^2.
\end{aligned}$$

Therefore, the Bernstein inequality implies that for all $x > 0$, we have

$$\mathbb{P}\Big(\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^*, \mathbf{E}\rangle \geqslant \sqrt{2x\rho_{\boldsymbol{\Sigma}}}\, \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathsf{F}} + (4/3)\rho_{\boldsymbol{\Theta}} x\Big) \leqslant e^{-x}.$$

Let us define $\Omega_{\mathbf{Z},\mathbf{Z}'} = \{(\widehat{\mathbf{Z}}^{\mathrm{user}}, \widehat{\mathbf{Z}}^{\mathrm{item}}) = (\mathbf{Z}, \mathbf{Z}')\}$, for each pair of matrices $\mathbf{Z} \in \mathcal{Z}_{n,K,n_0}$ and $\mathbf{Z}' \in \mathcal{Z}_{m,L,m_0}$. On the event $\Omega_{\mathbf{Z},\mathbf{Z}'}$, the matrix $\Pi_{\widehat{\mathcal{T}}}\boldsymbol{\Theta}^*$ is deterministic and its elements are averages of the elements of $\boldsymbol{\Theta}^*$. As $\Pi_{\widehat{\mathcal{T}}}$ is an orthogonal projection, we have $\|\Pi_{\widehat{\mathcal{T}}}\boldsymbol{\Theta}^*\|_{1,\infty} \leqslant \|\boldsymbol{\Theta}^*\|_{1,\infty} \leq \rho_{\boldsymbol{\Theta}}$ and

$$\mathbb{P}\Big(\big\{\langle \Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*, \mathbf{E}\rangle \geqslant \sqrt{2x\rho_{\boldsymbol{\Sigma}}}\, \|\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*\|_{\mathsf{F}} + (4/3)\rho_{\boldsymbol{\Theta}} x\big\} \cap \Omega_{\mathbf{Z},\mathbf{Z}'}\Big) \leq e^{-x}.$$

Note also that the cardinality of $\mathcal{Z}_{n,K,n_0}$ is at most K^n. Combining the last display with the union bound, we get

$$\begin{aligned}
&\mathbb{P}\Big(\langle \Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*, \mathbf{E}\rangle \geqslant \sqrt{2x\rho_{\boldsymbol{\Sigma}}}\, \|\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*\|_{\mathsf{F}} + (4/3)\rho_{\boldsymbol{\Theta}} x\Big)\\
&\qquad \leqslant \sum_{(\mathbf{Z},\mathbf{Z}')} \mathbb{P}\Big(\big\{\langle \Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*, \mathbf{E}\rangle \geqslant \sqrt{2x\rho_{\boldsymbol{\Sigma}}}\, \|\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*\|_{\mathsf{F}} + (4/3)\rho_{\boldsymbol{\Theta}} x\big\} \cap \Omega_{\mathbf{Z},\mathbf{Z}'}\Big)\\
&\qquad \leqslant K^n L^m e^{-x},
\end{aligned}$$

where in the first inequality in the above display the sum is over $(\mathbf{Z}, \mathbf{Z}')$ from the set $\mathcal{Z}_{n,K,n_0} \times \mathcal{Z}_{m,L,m_0}$ and the factor $K^n L^m$ corresponds to an upper bound on the cardinality of this set. Finally, choosing $x = n\log K + m\log L + t$ for some $t > 0$ and using the basic inequality $uv \leqslant \lambda u^2 + v^2/(4\lambda)$ entails

$$\mathbb{P}\left(\Xi_1 \geqslant \lambda \|\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2 + \Big(\frac{\rho_{\boldsymbol{\Sigma}}}{2\lambda} + \frac{4\rho_{\boldsymbol{\Theta}}}{3}\Big)(n\log K + m\log L + t)\right) \leqslant e^{-t}$$

for any $\lambda > 0$. Lemma 21 ensures that

$$\begin{aligned}
\mathbb{E}(\Xi_1) &\leqslant \lambda \mathbb{E}\Big[\|\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}^*) - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\Big] + \Big(\frac{\rho_{\boldsymbol{\Sigma}}}{2\lambda} + \frac{4\rho_{\boldsymbol{\Theta}}}{3}\Big)(n\log K + m\log L + 1)\\
&\leqslant \lambda \mathbb{E}\Big[\|\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2\Big] + \Big(\frac{\rho_{\boldsymbol{\Sigma}}}{2\lambda} + \frac{4\rho_{\boldsymbol{\Theta}}}{3}\Big)(n\log K + m\log L + 1).
\end{aligned}$$

Optimizing with respect to $\lambda > 0$, we get

$$\mathbb{E}(\Xi_1) \leqslant \sqrt{2\rho_{\mathbf{\Sigma}}(n\log K + m\log L + 1)}\,\mathbb{E}[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2} + (4/3)\rho_{\mathbf{\Theta}}(n\log K + m\log L + 1).$$

This completes the proof of the lemma. ■

We now switch to the evaluation of $\mathbb{E}(\Xi_2)$. To this end, we first notice that

$$\mathbb{E}(\Xi_2) = \mathbb{E}\big[\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{H}) - \Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta}^*), \mathbf{E}\rangle\big] = \mathbb{E}\big[\langle \Pi_{\widehat{\mathcal{T}}}(\mathbf{E}), \mathbf{E}\rangle\big] = \mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big].$$

Lemma 18. *Under the conditions of Proposition 5, we have*

$$\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big] \leqslant \Big(2\rho_{\mathbf{\Sigma}} + 2(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})^2\psi_{n,m}(n_0, m_0)\Big)(3KL + n\log K + m\log L).$$

Proof. The idea is to apply Lemma 24 and Lemma 25 together. We denote by $\mathcal{G}$ the family of all partitions $\mathcal{T}_0 = \{B_k \times C_\ell : k \in [K], \ell \in [L]\}$ of $[n] \times [m]$ such that for each $k \in [n]$, and $\ell \in [m]$, $|B_k| \geqslant n_0$ and $|C_\ell| \geqslant m_0$. We have

$$\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2 \leqslant \max_{\mathcal{T}_0 \in \mathcal{G}} \|\Pi_{\mathcal{T}_0}\mathbf{E}\|_{\mathsf{F}}^2$$

with rows of $\mathbf{E}$ satisfying the assumptions of Lemma 24 ($\|\mathbf{E}\|_{1,\infty} \leqslant \rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}}$). So by the union bound on $\mathcal{G}$, we obtain that

$$\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2 \leqslant 8(t + \log(2M|\mathcal{G}|))\big(\rho_{\mathbf{\Sigma}} + (2/3)(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})\max_{\mathcal{T}_0 \in \mathcal{G}} \|\Pi_{\mathcal{T}_0}\mathbf{E}\|_\infty\big) \tag{3.11}$$

with probability at least $1 - 0.5e^{-t}$, where $\log M \leqslant KL\log 12 \leqslant 2.5KL$. According to Lemma 25, on an event of probability at least $1 - 0.5e^{-t}$

$$\max_{\mathcal{T}_0 \in \mathcal{G}} \|\Pi_{\mathcal{T}_0}\mathbf{E}\|_\infty \leqslant \max_{B,C}\left|\sum_{i\in B} \frac{\boldsymbol{E}_i^\top \mathbf{1}_C}{|B||C|}\right| \leqslant \sqrt{\frac{2\rho_{\mathbf{\Sigma}}}{n_0 m_0}(t + \mathfrak{F})} + \frac{2(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})}{3n_0 m_0}(t + \mathfrak{F}),$$

where $\mathfrak{F} = \frac{\log(4nm)}{n_0 m_0} + \psi_{n,m}(n_0, m_0) \geqslant \frac{5.5}{n_0 m_0}$ if $n, m \geqslant 3$, and the maximum is taken over all pairs of subsets $B \subset [n]$ and $C \subset [m]$ such that $B \times C \in \mathcal{T}_0$. Hence with probability at least $1 - 0.5e^{-t}$

$$\begin{aligned}\max_{\mathcal{T}_0 \in \mathcal{G}} \|\Pi_G \vec{\mathbf{E}}\|_\infty &\leqslant \sqrt{2\rho_{\mathbf{\Sigma}}\mathfrak{F}(2t/11 + 1)} + 2/3(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})\mathfrak{F}(2t/11 + 1) \\ &\leqslant \sqrt{2\rho_{\mathbf{\Sigma}}\mathfrak{F}}(t/11 + 1) + 2/3(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})\mathfrak{F}(2t/11 + 1)\end{aligned} \tag{3.12}$$

Finally, combining (3.11) and (3.12), we have that with probability at least $1 - e^{-t}$,

$$\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2 \leqslant \big(t + \log(2M|\mathcal{G}|)\big)\Big(\rho_{\mathbf{\Sigma}} + 2/3(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})\sqrt{2\rho_{\mathbf{\Sigma}}\mathfrak{F}}(t/11 + 1)$$

$$+4/9(\rho_{\mathbf{H}}+\rho_{\mathbf{\Theta}})^2\mathfrak{F}(2t/11+1)\Big).$$

Using Lemma 21, we obtain the following upper bound on the expectation

$$\begin{aligned}\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big] &\leqslant \big(1+\log(2M|\mathcal{G}|)\big)\Big(\rho_{\mathbf{\Sigma}}+2/3(\rho_{\mathbf{H}}+\rho_{\mathbf{\Theta}})\sqrt{2\rho_{\mathbf{\Sigma}}\mathfrak{F}}(2/11+1)\\ &\qquad +4/9(\rho_{\mathbf{H}}+\rho_{\mathbf{\Theta}})^2\mathfrak{F}(4/11+1)\Big)\\ &\leqslant \big(1+\log(2M|\mathcal{G}|)\big)\Big(2\rho_{\mathbf{\Sigma}}+(\rho_{\mathbf{H}}+\rho_{\mathbf{\Theta}})^2\mathfrak{F}\Big)\end{aligned}$$

But $|\mathcal{G}|\leqslant K^nL^m$ and $M\leqslant 12^{KL}$. Taking into account the fact that $K\geqslant 2$ and $L\geqslant 2$, this leads to

$$\begin{aligned}1+\log(2M|\mathcal{G}|) &\leqslant 1+\log 2+KL\log 12+n\log K+m\log L\\ &\leqslant 3KL+n\log K+m\log L.\end{aligned}$$

We also have that

$$\mathfrak{F}=\frac{\log(4nm)}{n_0m_0}+\psi_{n,m}(n_0,m_0)\leqslant 2\psi_{n,m}(n_0,m_0),$$

so the control of $\mathbf{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big]$ is finally given by

$$\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big]\leqslant\Big(2\rho_{\mathbf{\Sigma}}+2(\rho_{\mathbf{H}}+\rho_{\mathbf{\Theta}})^2\psi_{n,m}(n_0,m_0)\Big)(3KL+n\log K+m\log L).$$

This completes the proof of the lemma. ■

In order to ease notation in the rest of the proof, let us set $A=n\log K+m\log L+1$. To conclude, we use the bounds on Ξ_1 and Ξ_2 obtained in Lemma 17 and Lemma 18, respectively, as well as decompositions (3.9) and (3.10). This implies that

$$\begin{aligned}\mathbb{E}[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}}-\mathbf{\Theta}^*\|_{\mathsf{F}}^2] &\leqslant \|\mathbf{\Theta}^*-\Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\|_{\mathsf{F}}^2+2\Big(2\rho_{\mathbf{\Sigma}}A\,\mathbb{E}[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}}-\mathbf{\Theta}^*\|_{\mathsf{F}}^2]\Big)^{1/2}\\ &\qquad +(8/3)\rho_{\mathbf{\Theta}}A+2\mathbf{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big].\end{aligned}$$

One can check that the last inequality is equivalent to

$$\Big(\mathbb{E}[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}}-\mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}-\sqrt{2\rho_{\mathbf{\Sigma}}A}\Big)^2\leqslant\|\mathbf{\Theta}^*-\Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\|_{\mathsf{F}}^2+(2\rho_{\mathbf{\Sigma}}+3\rho_{\mathbf{\Theta}})A+2\mathbf{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big].$$

This readily yields

$$\begin{aligned}&\mathbb{E}[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}}-\mathbf{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}\\ &\quad\leqslant\|\mathbf{\Theta}^*-\Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\|_{\mathsf{F}}+\sqrt{2\rho_{\mathbf{\Sigma}}A}+\big((2\rho_{\mathbf{\Sigma}}+3\rho_{\mathbf{\Theta}})A+2\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big]\big)^{1/2}\\ &\quad\leqslant\|\mathbf{\Theta}^*-\Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\|_{\mathsf{F}}+\big((8\rho_{\mathbf{\Sigma}}+6\rho_{\mathbf{\Theta}})A+4\mathbb{E}\big[\|\Pi_{\widehat{\mathcal{T}}}(\mathbf{E})\|_{\mathsf{F}}^2\big]\big)^{1/2}\end{aligned}$$

$$\leqslant \|\mathbf{\Theta}^* - \Pi_{\mathcal{T}}(\mathbf{\Theta}^*)\|_{\mathsf{F}} + \left(16\rho_{\mathbf{\Sigma}} + 6\rho_{\mathbf{\Theta}} + 8(\rho_{\mathbf{H}} + \rho_{\mathbf{\Theta}})^2\psi_{n,m}(n_0, m_0)\right)^{1/2} r_{n,m}(K, L),$$

where in the second line we have used that $\sqrt{a} + \sqrt{b} \leqslant \sqrt{2a + 2b}$ and we also denoted $r_{n,m}(K, L) = \left(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\right)^{1/2}$.

3.4.2 Proof of Proposition 6 (identifiability property)

Proof. It is easy to see that (3.5) is equivalent to

$$\sum_{\ell=1}^{m} W(x, y_1)W'(\tau(x), \tau_\ell(y_\ell)) = \sum_{\ell=1}^{m} W'(\tau(x), \tau_1(y_1))W(x, y_\ell)\,.$$

Since the first term of each sum is identical, we can factor it out of the equation. Then integrated this equation with respect to $y_1, \dots y_{m-1}$, we get

$$\begin{aligned}(m-2)I_W(x)I_{W'}(\tau(x)) + I_W(x)W'(\tau(x), \tau_m(y_m)) \\ = (m-2)I_W(x)I_{W'}(\tau(x)) + I_{W'}(\tau(x))W(x, y_m)\end{aligned}$$

where $I_W(x) = \displaystyle\int_0^1 W(x, y)dy$ and $I_{W'}(x) = \displaystyle\int_0^1 W'(x, y)dy$. Finally for all $x, y \in [0, 1]$,

$$W(x, y) = \nu(x)W'(\tau(x), \tau_m(y)).$$

with $\nu(x) = \dfrac{I_W(x)}{I_{W'}(\tau(x))}$. □

3.4.3 Proof of Lemma 16 ($\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ is left-stochastic)

Recall that $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \Pi_{\widehat{\mathcal{T}}}\mathbf{H}$ with $\widehat{\mathcal{T}} = \left\{\mathbf{\Theta} : \exists \mathbf{Q} \in [0, 1]^{K\times L} \text{ such that } \mathbf{\Theta} = \widehat{\mathbf{Z}}^{\text{user}}\mathbf{Q}(\widehat{\mathbf{Z}}^{\text{item}})^\top\right\}$ the linear space of constant by block matrices, where the blocks are the same than $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$. The proof consists in given an exact expression of $\Pi_{\widehat{\mathcal{T}}}(\mathbf{\Theta})$ for any matrix $\mathbf{\Theta} \in [0, 1]^{n\times m}$ with rows summing to one. Let define $\mathbf{\Theta}_1 = \widehat{\mathbf{Z}}^{\text{user}}\mathbf{Q}_1(\widehat{\mathbf{Z}}^{\text{item}})^\top \in \widehat{\mathcal{T}}$ where $\mathbf{Q}_1 \in \mathbb{R}^{K\times L}$ has as entries some averages of entries of $\mathbf{\Theta}$, or more precisely

$$(Q_1)_{k\ell} = \frac{1}{|\widehat{B}_k||\widehat{C}_\ell|}\sum_{i\in\widehat{B}_k}\sum_{j\in\widehat{C}_\ell}\Theta_{ij}\,.$$

Then, for all $\mathbf{\Theta}' = \widehat{\mathbf{Z}}^{\text{user}}\mathbf{Q}'(\widehat{\mathbf{Z}}^{\text{item}})^\top \in \widehat{\mathcal{T}}$,

$$\|\mathbf{\Theta}' - \mathbf{\Theta}\|_{\mathsf{F}}^2 = \|\mathbf{\Theta}' - \mathbf{\Theta}_1\|_{\mathsf{F}}^2 + \|\mathbf{\Theta}_1 - \mathbf{\Theta}\|_{\mathsf{F}}^2 + 2\langle\mathbf{\Theta}' - \mathbf{\Theta}_1, \mathbf{\Theta}_1 - \mathbf{\Theta}\rangle\,. \tag{3.13}$$

but

$$\begin{aligned}\langle \boldsymbol{\Theta}' - \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_1 - \boldsymbol{\Theta} \rangle &= \sum_{k=1}^{K}\sum_{\ell=1}^{L}\sum_{i\in\widehat{B}_k}\sum_{j\in\widehat{C}_\ell}\big(Q'_{k\ell} - (Q_1)_{k\ell}\big)\big((Q_1)_{k\ell} - \Theta_{ij}\big) \\ &= \sum_{k=1}^{K}\sum_{\ell=1}^{L}\big(Q'_{k\ell} - (Q_1)_{k\ell}\big)\underbrace{\sum_{i\in\widehat{B}_k}\sum_{j\in\widehat{C}_\ell}\big((Q_1)_{k\ell} - \Theta_{ij}\big)}_{=0} \\ &= 0\end{aligned}$$

Then according to (3.13), $\boldsymbol{\Theta}_1$ is the solution to

$$\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}) = \arg\min_{\boldsymbol{\Theta}'\in\widehat{\mathcal{T}}} \|\boldsymbol{\Theta}' - \boldsymbol{\Theta}\|_{\mathsf{F}}.$$

One can easily check that every rows of $\boldsymbol{\Theta}_1$ sum to one, and we have proved that $\Pi_{\widehat{\mathcal{T}}}(\boldsymbol{\Theta}) = \boldsymbol{\Theta}_1$. Taking $\boldsymbol{\Theta} = \mathbf{H}$ above prove the lemma.

3.4.4 Proof of Proposition 7 (approximation error for a graphon)

In what follows, λ refers to the Lebesgue measure on $\mathbb{R}$ and λ_2 is the Lebesgue measure on $\mathbb{R}^2$. Let W^* be a graphon such that for some $K \times L$ matrix $\mathbf{Q}^*$ and some sequences $a_0 < \ldots < a_K$, $b_0 < \ldots < b_L$ satisfying $a_0 = b_0 = 0$ and $a_K = b_L = 1$, we have $W^*(u,v) = Q^*_{k,\ell}$ for every $u \in [a_{k-1}, a_k)$ and $v \in [b_{\ell-1}, b_\ell)$. Equivalently,

$$W^*(u,v) = \sum_{k=1}^{K}\sum_{\ell=1}^{L} Q^*_{k,\ell}\mathbb{1}_{[a_{k-1},a_k)\times[b_{\ell-1},b_\ell)}(u,v).$$

The condition $W^* \in \mathscr{C}$ can be rewritten as

$$\sum_{\ell=1}^{L} Q^*_{k,\ell} w^{(2)}_\ell = \frac{1}{m} \qquad \forall k \in [K]$$

where we define the "weight" sequences $w^{(1)}_k = a_k - a_{k-1}$, $w^{(2)}_\ell = b_\ell - b_{\ell-1}$ and

$$\widehat{w}^{(1)}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[a_{k-1},a_k[}(U_i) \qquad \text{and} \qquad \widehat{w}^{(2)}_\ell = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}_{[b_{\ell-1},b_\ell[}(V_j).$$

Notice that all the four weight sequences $w^{(1)}$, $w^{(2)}$, $\widehat{w}^{(1)}$ and $\widehat{w}^{(2)}$ are positive and sum to one. As proved in [KTV17, p16], there exist two functions $\psi_1 : [0,1] \to [K]$ and $\psi_2 : [0,1] \to [L]$ such that

1. For all $k \in [K]$ and $x \in \big[a_{k-1}, (a_{k-1} + \widehat{w}^{(1)}_k) \wedge a_k\big)$, we have $\psi_1(x) = k$
2. For all $\ell \in [L]$ and $x \in \big[b_{\ell-1}, (b_{\ell-1} + \widehat{w}^{(2)}_\ell) \wedge b_\ell\big)$, we have $\psi_2(x) = \ell$

3. $\lambda(\psi_1^{-1}(k)) = \widehat{w}_k^{(1)}$ for all $k \in [K]$
4. $\lambda(\psi_2^{-1}(\ell)) = \widehat{w}_\ell^{(2)}$ for all $\ell \in [L]$.

Using these mappings ψ_1 and ψ_2, we construct the graphon $W_\psi^*(u,v) = \mathbf{Q}^*_{\psi_1(u),\psi_2(v)}$.

Lemma 19. *W_ψ^* satisfies $\delta(W_{\mathbf{\Theta}^*}, (mI_{W_{\psi^*}})^{-1}W_\psi^*) = 0$.*

Proof. Indeed, it suffices to prove that there exist two measure preserving automorphisms of $[0,1]$ called τ_1 and τ_2 such that $W_\psi^*(u,v) = \nu(u)W_{\mathbf{\Theta}^*}(\tau_1(u),\tau_2(v))$ for almost all $u,v \in [0,1]$ where $\nu(u)$ is a positive random variable that could depend on u. On the one hand, one can easily check that for any $k \in [K]$

$$\lambda\Big(\big\{u,\ \ W_\psi^*(u,v) = \mathbf{Q}^*_{k,\ell} \text{ for some } \ell \in [L]\big\}\Big) = \lambda\big(\{u,\ \ \psi_1(u) = k\}\big) = \widehat{w}_k^{(1)}.$$

On the other hand, we have

$$\begin{aligned} W_{\mathbf{\Theta}^*}(u,v) &= \mathbf{\Theta}^*_{\lceil nu\rceil,\lceil mv\rceil} \\ &= \frac{W^*(U_{\lceil nu\rceil}, V_{\lceil mv\rceil})}{\sum_{j=1}^m W^*(U_{\lceil nu\rceil}, V_j)} \\ &= \nu(u)W^*(U_{\lceil nu\rceil}, V_{\lceil mv\rceil}). \end{aligned}$$

But

$$\begin{aligned} \lambda\Big(\Big\{u,\ \ & W^*(U_{\lceil nu\rceil}, V_{\lceil mv\rceil}) \\ & = \mathbf{Q}^*_{k,\ell} \text{ for some } \ell \in [L]\Big\}\Big) = \lambda\big(\{u,\ \ U_{\lceil nu\rceil} \in [a_{k-1}, a_k)\}\big) \\ &= \lambda\Big(\bigcup_{i=1}^n \Big\{u,\ \ \frac{i-1}{n} \leqslant u < \frac{i}{n} \text{ and } U_i \in [a_{k-1}, a_k)\Big\}\Big) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i \in [a_{k-1}, a_k)} \\ &= \widehat{w}_k^{(1)}. \end{aligned}$$

Then for any $k \in [K]$ we can find a measure preserving bijection π_k from the interior of $\psi_1^{-1}(k)$ to the interior of $\big\{u,\ \ U_{\lceil nu\rceil} \in [a_{k-1}, a_k)\big\}$. Notice that each of these subsets is a finite reunion of intervals. Lemma 22 prove the existence of π_k. Now we construct τ_1 such that $\tau_1|_{\overset{\circ}{\psi_1^{-1}(k)}} = \pi_k$ which is a measure preserving automorphism of $[0,1]$. We have a similar outcome for the v-axis, which gives τ_2, and proves the lemma. ■

This lemma and the triangular inequality lead to

$$\delta(W_{\mathbf{\Theta}^*}, W^*) = \delta((mI_{W_\psi^*})^{-1}W_\psi^*, W^*) \leqslant \|(mI_{W_\psi^*})^{-1}W_\psi^* - W_\psi^*\|_{\mathbb{L}_2} + \|W_\psi^* - W^*\|_{\mathbb{L}_2}.$$

The second term is upper bounded in expectation by $\frac{(B-A)}{\sqrt{2}}\left(\sqrt{\frac{K}{n}}+\sqrt{\frac{L}{m}}\right)^{1/2}$ (see Subsection 2.7.2). It now remains to upper-bound the expectation of first term.

$$\begin{aligned}\|(mI_{W_\psi^*})^{-1}W_\psi^* - W_\psi^*\|_{\mathbb{L}_2}^2 &= \int_{[0,1]^2} W_\psi^*(u,v)^2\Big(1-(mI_{W_\psi^*})^{-1}(u)\Big)^2 dudv \\ &= \sum_{k=1}^K\sum_{\ell=1}^L \widehat{w}_k^{(1)}\widehat{w}_\ell^{(2)}(Q_{k\ell}^*)^2\Big(\frac{1-m\sum_{\ell'=1}^L Q_{k\ell'}^*\widehat{w}_{\ell'}^{(2)}}{m\sum_{\ell'=1}^L Q_{k\ell'}^*\widehat{w}_{\ell'}^{(2)}}\Big)^2 \\ &= \sum_{k=1}^K \widehat{w}_k^{(1)}\frac{\sum_{\ell=1}^L (Q_{k\ell}^*)^2\widehat{w}_\ell^{(2)}}{\Big(\sum_{\ell'=1}^L Q_{k\ell'}^*\widehat{w}_{\ell'}^{(2)}\Big)^2}\Big(\frac{1}{m}-\sum_{\ell'=1}^L Q_{k\ell'}^*\widehat{w}_{\ell'}^{(2)}\Big)^2.\end{aligned}$$

Using Lemma 20 bellow we get

$$\mathbb{E}[\|(mI_{W_\psi^*})^{-1}W_\psi^* - W_\psi^*\|_{\mathbb{L}_2}^2] \leqslant \frac{3B^2}{m}\,.$$

This ends the proof of Proposition 7.

Lemma 20. *Let $V_1,\dots,V_m$ iid random variable drawn from the uniform distribution on $[0,1]$. Let $C_1,\dots,C_L$ be a partition of $[0,1]$ with $w_\ell=\lambda(C_\ell)$. We set*

$$Z_{\ell j} = \mathbb{1}_{V_j\in C_\ell} \quad \textit{and} \quad \widehat{w}_\ell = \frac{1}{m}\sum_{j=1}^m Z_{\ell j}.$$

for every sequence $(a_1,\dots,a_L)$ of real number such that

$$0\leqslant a_\ell \leqslant B \;\; \forall \ell\in[L] \quad \textit{and} \quad \sum_{\ell=1}^L a_\ell w_\ell = \frac{1}{m},$$

denoting $\xi = \dfrac{\sum_{\ell=1}^L a_\ell^2\widehat{w}_\ell}{\Big(\sum_{\ell=1}^L a_\ell\widehat{w}_\ell\Big)^2}\Big(\dfrac{1}{m}-\sum_{\ell=1}^L a_\ell\widehat{w}_\ell\Big)^2$, we have for every $t\in(0,1)$

$$\mathbb{E}[\xi] \leqslant \frac{B^2}{tm} + \frac{4B^2}{m}\exp\left(\frac{-3m(1-t)^2}{16B}\right)\bar{w},$$

where $\bar{w}=\sum_{\ell=1}^L\frac{1}{w_\ell}$. Moreover, if $\bar{w}\leqslant\frac{1}{4}\exp\left(\frac{3}{64}m\right)$, then

$$\mathbb{E}[\xi] \leqslant \frac{3B^2}{m}.$$

Proof. For some real number $t>0$ let denote $\Omega_{>t} = \Big\{\sum_{\ell=1}^L a_\ell\widehat{w}_\ell > t\Big\}$. Then $\mathbb{E}[\xi] =$

$\mathbb{E}[\xi \mathbb{1}_{\Omega_{>t}}] + \mathbb{E}[\xi \mathbb{1}_{\Omega_{\leqslant t}}]$ and the aim is now to upper-bound each term. On the one hand

$$\begin{aligned}\mathbb{E}[\xi \mathbb{1}_{\Omega_{>t}}] &\leqslant \frac{B}{t}\mathbb{E}\Big[\Big(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\Big)^2\Big] \\ &= \frac{B}{t}\mathbf{Var}\Big[\frac{1}{m}\sum_{j=1}^{m} Y_j\Big]\end{aligned}$$

where $Y_j = \sum_{\ell=1}^{L} a_\ell Z_{\ell j}$ are iid random variables. We have

$$\mathbf{Var}(Y_j) = \mathbb{E}[Y_j^2] - 1 = \sum_{\ell=1}^{L} a_\ell^2 w_\ell - 1 \leqslant B - 1,$$

which leads to

$$\mathbb{E}[\xi \mathbb{1}_{\Omega_{>t}}] \leqslant \frac{B^2}{tm}.$$

On the other hand

$$\begin{aligned}\mathbb{E}[\xi \mathbb{1}_{\Omega_{\leqslant t}}] &= \mathbb{E}[\xi \mathbb{1}_{\Omega_{\leqslant t}} \mathbb{1}_{\Omega_{>0}}] \\ &= \mathbb{E}\left[\frac{\sum_{\ell=1}^{L} a_\ell^2 \widehat{w}_\ell}{\Big(\sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\Big)^2}\Big(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\Big)^2 \mathbb{1}_{\Omega_{\leqslant t}} \mathbb{1}_{\Omega_{>0}}\right] \\ &= \mathbb{E}\left[\sum_{\ell=1}^{L} \frac{a_\ell^2 \widehat{w}_\ell \mathbb{1}_{\widehat{w}_\ell > 0}}{\Big(\sum_{\ell'=1}^{L} a_{\ell'} \widehat{w}_{\ell'}\Big)^2}\Big(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\Big)^2 \mathbb{1}_{\Omega_{\leqslant t}} \mathbb{1}_{\Omega_{>0}}\right] \\ &\leqslant \mathbb{E}\left[\sum_{\ell=1}^{L} \frac{\mathbb{1}_{\widehat{w}_\ell > 0}}{\widehat{w}_\ell}\Big(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\Big)^2 \mathbb{1}_{\Omega_{\leqslant t}}\right] \\ &\leqslant \mathbb{E}\left[\sum_{\ell=1}^{L} \frac{\mathbb{1}_{\widehat{w}_\ell > 0}}{\widehat{w}_\ell^2}\right]^{1/2} \mathbb{E}\left[\Big(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\Big)^4 \mathbb{1}_{\Omega_{\leqslant t}}\right]^{1/2} \\ &\leqslant \mathbb{E}\left[\Big(\sum_{\ell=1}^{L} \frac{\mathbb{1}_{\widehat{w}_\ell > 0}}{\widehat{w}_\ell}\Big)^2\right]^{1/2} \mathbb{E}\left[\Big(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\Big)^8\right]^{1/4} \mathbb{P}(\mathbb{1}_{\Omega_{\leqslant t}})^{1/4}\end{aligned}$$

where we used multiple times the Cauchy-Schwartz inequality. Now it remains to upper-bound each factor individually. For the third factor, assume that $t \in (0,1)$, we use the classical Bernstein inequality

$$\begin{aligned}\mathbb{P}(\Omega_{\leqslant t}) &= \mathbb{P}\Big(\sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell \leqslant t\Big) \\ &= \mathbb{P}\Big(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell \geqslant 1 - t)\Big)\end{aligned}$$

$$= \mathbb{P}\Big(\mathbb{E}[Y_j] - \frac{1}{m}\sum_{j=1}^{m} Y_j \geqslant 1 - t\Big).$$

Recalling that $Y_j = \sum_{\ell=1}^{L} a_\ell Z_{\ell j}$ are iid random variables, satisfying $|Y_j| \leqslant B$ and $\mathbb{E}[Y_j^2] \leqslant B$, the Bernstein inequality gives

$$\mathbb{P}(\Omega_{\leqslant t}) \leqslant \exp\left(\frac{-m(1-t)^2}{B(1+\frac{1-t}{3})}\right) \leqslant \exp\left(\frac{-3m(1-t)^2}{4B}\right).$$

We can use once again the Hoeffding inequality and integration by parts arguments for the second factor. Indeed

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{m} - \sum_{\ell=1}^{L} a_\ell \widehat{w}_\ell\right)^8\right] &= \mathbb{E}[Y^8] \\
&= 8\int_0^\infty x^7 \mathbb{P}(|Y| > x)dx \\
&\leqslant 16\int_0^\infty x^7 \exp\left(\frac{-2mx^2}{B^2}\right)dx \\
&= 3\Big(\frac{B^2}{m}\Big)^4.
\end{aligned}$$

Finally, we use the trangular inequality together with Lemma 23 to get a bound on the first term

$$\begin{aligned}
\mathbb{E}\left[\Big(\sum_{\ell=1}^{L} \frac{\mathbb{1}_{\widehat{w}_\ell > 0}}{\widehat{w}_\ell}\Big)^2\right]^{1/2} &\leqslant \sum_{\ell=1}^{L} \mathbb{E}\Big[\frac{\mathbb{1}_{\widehat{w}_\ell>0}}{\widehat{w}_\ell^2}\Big]^{1/2} \\
&\leqslant 2\sqrt{2}\sum_{\ell=1}^{L}\frac{1}{w_\ell}.
\end{aligned}$$

For the final claim, we choose $t = 1 - \sqrt{\frac{16}{3m}\log(4\bar{w})} \geqslant \frac{1}{2}$ provided that $\bar{w} \leqslant \frac{1}{4}e^{0.045m}$. ■

3.4.5 Proof of Theorem 10 (risk bound for $\widehat{W}^{\mathsf{LS}}$)

In view of (3.7), the fact that $\widehat{W} = W_{\widehat{\boldsymbol{\Theta}}}$ and Proposition 7, we have

$$\begin{aligned}
\mathbb{E}[\delta(W_{\widehat{\boldsymbol{\Theta}}}, W^*)^2]^{1/2} &\leqslant \frac{\mathbb{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}}{\sqrt{nm}} + \mathbb{E}[\delta(W_{\boldsymbol{\Theta}^*}, W^*)^2]^{1/2} \\
&\leqslant \frac{\mathbb{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}}{\sqrt{nm}} + \frac{\rho}{\sqrt{2}}\left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\right)^{1/2} + \rho\sqrt{\frac{3}{m}} \\
&\leqslant \frac{\mathbb{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2]^{1/2}}{\sqrt{nm}} + 2\rho\left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\right)^{1/2}. \qquad (3.14)
\end{aligned}$$

Let $\widehat{\mathcal{T}}$ and $\mathcal{T}^*$ be the sets of all $n \times m$ matrices with real entries that are constant by block on the same blocks as $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^*$, respectively. Clearly, $\widehat{\mathcal{T}}$ and $\mathcal{T}^*$ are linear subspaces of the

space of $n \times m$ real matrices equipped with the scalar product $\langle \mathbf{M}_1, \mathbf{M}_2 \rangle = \mathrm{tr}(\mathbf{M}_1^\top \mathbf{M}_2)$. Let $\Pi_{\widehat{\mathcal{T}}}$ and $\Pi_{\mathcal{T}^*}$ be the orthogonal projections onto $\widehat{\mathcal{T}}$ and $\mathcal{T}^*$ respectively. We have $\Pi_{\widehat{\mathcal{T}}} \mathbf{H} = \widehat{\mathbf{\Theta}}$ and $\Pi_{\mathcal{T}^*} \mathbf{\Theta}^* = \mathbf{\Theta}^*$. Therefore,

$$\begin{aligned}
\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}} &= \|\Pi_{\widehat{\mathcal{T}}} \mathbf{H} - \mathbf{\Theta}^*\|_{\mathsf{F}} \\
&\overset{①}{\leqslant} \|\Pi_{\widehat{\mathcal{T}}}(\mathbf{H} - \mathbf{\Theta}^*)\|_{\mathsf{F}} + \|\Pi_{\widehat{\mathcal{T}}} \mathbf{\Theta}^* - \mathbf{\Theta}^*\|_{\mathsf{F}} \\
&\overset{②}{\leqslant} \|\mathbf{H} - \mathbf{\Theta}^*\|_{\mathsf{F}} + \|(\rho/2)\mathbf{1}_n \mathbf{1}_m^\top - \mathbf{\Theta}^*\|_{\mathsf{F}}.
\end{aligned}$$

Above, ① is a consequence of the triangle inequality, whereas ② follows from the fact that $\Pi_{\widehat{\mathcal{T}}}$ is an orthogonal projection (hence, a contraction) and the matrix $(\rho/2)\mathbf{1}_n \mathbf{1}_m^\top$ belongs to the image of $\Pi_{\widehat{\mathcal{T}}}$. Hence

$$\frac{1}{nm} \mathbb{E}\big[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \,\big|\, \boldsymbol{U}, \boldsymbol{V}\big] \leqslant (\sqrt{\rho_{\mathbf{\Sigma}}} + 0.5\rho)^2.$$

For every $k \in [K]$ and $\ell \in [L]$, we define $n_k = n|a_k - a_{k-1}|$, $N_k = \#\{i : U_i \in [a_{k-1}, a_k[\}$, $m_\ell = m|b_\ell - b_{\ell-1}|$ and $M_\ell = \#\{j : V_j \in [b_{\ell-1}, b_\ell[\}$. We also define the event $\Omega_0 = \{N_k \geqslant n_k/2; M_\ell \geqslant m_\ell/2 \text{ for all } k \in [K] \text{ and } \ell \in [L]\}$. Since the event Ω_0^c is $(\boldsymbol{U}, \boldsymbol{V})$-measurable, we get

$$\frac{1}{nm} \mathbb{E}\big[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \mathbb{1}_{\Omega_0^c}\big] = \frac{1}{nm} \mathbb{E}\Big(\mathbb{E}\big[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \,\big|\, \boldsymbol{U}, \boldsymbol{V}\big] \mathbb{1}_{\Omega_0^c}\Big) \leqslant (\sqrt{\rho_{\mathbf{\Sigma}}} + 0.5\rho)^2 \mathbb{P}(\Omega_0^c).$$

Using the union bound and the Chernoff inequality, one can check that

$$\mathbb{P}(\Omega_0^c) \leqslant \sum_{k=1}^K \mathbb{P}(N_k \leqslant n_k/2) + \sum_{\ell=1}^L \mathbb{P}(M_\ell \leqslant m_\ell/2) \leqslant \sum_{k=1}^K e^{-n_k/8} + \sum_{\ell=1}^L e^{-m_\ell/8}.$$

Since we have assumed that $n_k \geqslant 8\log(nK)$ and $m_\ell \geqslant 8\log(mL)$, we get $\mathbb{P}(\Omega_0^c) \leqslant n^{-1} + m^{-1}$. If the parameters n_0 and m_0 used in the definition of the least squares estimator $\widehat{\mathbf{\Theta}}$ satisfy $n_0 = \min_k n_k/2 = n\Delta^{(K)}/2$ and $m_0 = \min_\ell m_\ell/2 = m\Delta^{(L)}/2$, then on the event Ω_0 we ca apply Proposition 5. One can check that $\psi_{n,m}(n_0, m_0) = \psi_{n,m}(\Delta^{(K,L)})$. This, in conjunction with the previous inequalities, implies that

$$\begin{aligned}
\frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2]}{nm} &= \frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \mathbb{1}_{\Omega_0}]}{nm} + \frac{\mathbb{E}[\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2 \mathbb{1}_{\Omega_0^c}]}{nm} \\
&\leqslant \big(48\rho_{\mathbf{\Sigma}} + 6\rho\big)\Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big) + \frac{(\sqrt{\rho}_{\mathbf{\Sigma}} + 0.5\rho)^2}{n} + \frac{(\sqrt{\rho}_{\mathbf{\Sigma}} + 0.5\rho)^2}{m} \\
&\leqslant \bigg\{\big(50\rho_{\mathbf{\Sigma}} + 6\rho\big)^{1/2}\Big(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2} + \frac{\rho}{2}\sqrt{\frac{1}{n} + \frac{1}{m}}\bigg\}^2,
\end{aligned}$$

under condition that $\psi_{n,m}(\Delta^{(K,L)}) \leqslant \rho_{\mathbf{\Sigma}}$. One can also check that if $K, L \geqslant 2$ and $n, m \geqslant 5$, it holds

$$\frac{1}{n} + \frac{1}{m} \leq \frac{1}{3}\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big).$$

This inequality, combined with (3.14), completes the proof of the theorem.

3.5 Auxiliary results

Lemma 21. *Let X be a random variable and $a \in \mathbb{R}$, $b, c, d \geqslant 0$ be some constants. If*

$$\mathbb{P}(X \geqslant a + bt + ct^2) \leqslant de^{-t} \qquad \textit{for all } t \geqslant 0,$$

then $\mathbb{E}[X] \leqslant a + bd + 2cd$.

Proof. In the case $c = 0$, this inequality is well-known. Therefore, we consider only the case $c > 0$. Without loss of generality, we assume that $a = 0$ and $c = 1$. Indeed, we can always reduce to this case by considering the random variable $X' = (X - a)_+/c$ with $b' = b/c$. Thus, we know that $\mathbb{P}(X \geqslant t^2 + bt) \leqslant de^{-t}$ for every $t \geqslant 0$. Note that the condition $b \geqslant 0$ entails that the mapping $t \mapsto t^2 + bt$ defined on $[0, +\infty)$ is bijective. Setting $z = t^2 + bt$, this implies that

$$\mathbb{P}\big(X \geqslant z\big) \leqslant d \exp\big\{(b/2) - \sqrt{z + (b/2)^2}\big\}, \qquad \forall z \geq 0.$$

This inequality yields

$$\begin{aligned}
\mathbb{E}[X] &\leqslant d \int_0^\infty \exp\big\{(b/2) - \sqrt{z + (b/2)^2}\big\}\, dz \\
&= d \int_0^\infty e^{-t}(2t + b)\, dt \\
&= bd + 2d.
\end{aligned}$$

This completes the proof. □

Lemma 22. *Let $E \subset [0, 1]$ be a finite reunion of intervals with positive Lebesgue measure. Then there exists a measure preserving bijection $\pi : \mathring{E} \to [0, \lambda(E)[$.*

Proof. Let denote $E = \bigcup_{i=1}^n I_i$, with $I_i \cap I_j = \emptyset$. We define for every $x \in \mathring{E}$ (see Figure 20)

$$\pi(x) = \Big(x - \inf_y I_i + \sum_{j=1}^{i-1} \lambda(I_j)\Big)\mathbb{1}_{x \in I_i}$$

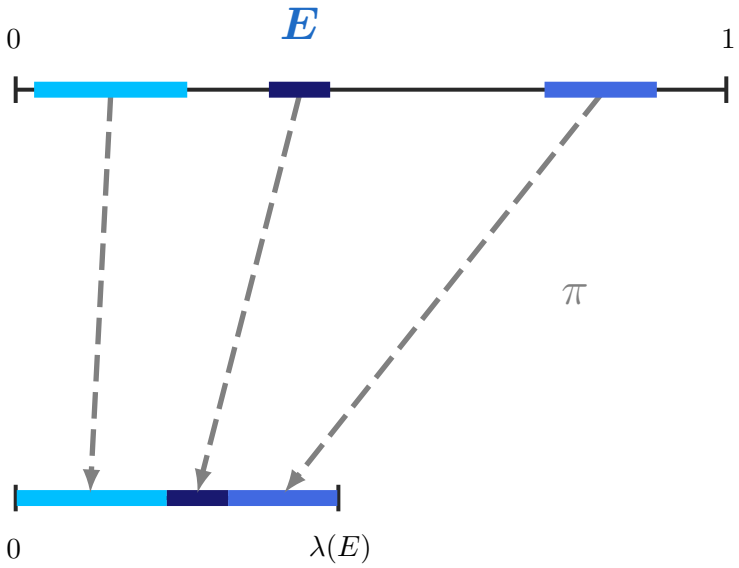


Figure 20: Measure preserving bijection from $\mathring{E}$ to $[0, \lambda(E)[$.

The map π is piecewise affine with directing coefficient 1, so it preserves the Lebesgue measure, and it is bijective. □

Lemma 23 (Upper-bounding the moments of the inverse of a binomial)**.** *Let X be a random variable with binomial distribution of parameter (n, p), we have for every $k \geqslant 1$*

$$\mathbb{E}\left[\frac{\mathbb{1}_{X>0}}{X^k}\right] \leqslant \frac{C_k}{(n+1)(n+2)\dots(n+k)p^k} \leqslant \frac{C_k}{(np)^k}.$$

with $C_k = 2^{\frac{k(k+1)}{2}}$

Proof. We prove the result by induction on $k \geqslant 1$. for the initialization, we use that $X\mathbb{1}_{X>0} \geqslant 1$ and $\frac{1}{t} \leqslant \frac{2}{t+1}$ for every $t \geqslant 1$. It gives

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbb{1}_{X>0}}{X}\right] &\leqslant \mathbb{E}\left[\frac{2}{X+1}\right] \\
&= \sum_{i=0}^{n} \frac{2}{i+1}\binom{n}{i} p^i (1-p)^{n-i} \\
&= \frac{2}{(n+1)p} \sum_{i=0}^{n} \binom{n+1}{i+1} p^{i+1} (1-p)^{n-i} \\
&\leqslant \frac{2}{(n+1)p} \sum_{j=0}^{n+1} \binom{n+1}{j} p^j (1-p)^{n+1-j} \\
&= \frac{2}{(n+1)p}.
\end{aligned}$$

Let assume that the statement is valid for some $k \geqslant 1$, then using the same trick than for the

initialization, we get

$$\begin{aligned}
\mathbb{E}\Big[\frac{\mathbb{1}_{X>0}}{X^{k+1}}\Big] &\leqslant \mathbb{E}\Big[\frac{2^{k+1}}{(X+1)^{k+1}}\Big] \\
&= \sum_{i=0}^{n} \frac{2^{k+1}}{(i+1)^{k+1}} \binom{n}{i} p^i (1-p)^{n-i} \\
&= \frac{2^{k+1}}{(n+1)p} \sum_{i=0}^{n} \frac{1}{(i+1)^k} \binom{n+1}{i+1} p^{i+1} (1-p)^{n-i} \\
&= \frac{2^{k+1}}{(n+1)p} \sum_{j=1}^{n+1} \frac{1}{j^k} \binom{n+1}{j} p^{j} (1-p)^{n+1-j} \\
&= \frac{2^{k+1}}{(n+1)p} \mathbb{E}\Big[\frac{\mathbb{1}_{Y>0}}{Y^k}\Big] \qquad Y \sim \mathcal{B}(n+1,p) \\
&\leqslant \frac{2^{k+1}}{(n+1)p} \frac{2^{\frac{k(k+1)}{2}}}{(n+2)(n+3)\dots(n+1+k)p^k}
\end{aligned}$$

The lemma is proved by induction on k. □

Lemma 24. *Let $\boldsymbol{E}_1, \dots, \boldsymbol{E}_n$ be centered and independent random vectors in $\mathbb{R}^m$ such that $\|\mathbb{E}[\boldsymbol{E}_i \boldsymbol{E}_i^\top]\|_{\mathrm{op}} \leqslant \rho_{\mathbf{\Sigma}}$ with $\rho_{\mathbf{\Sigma}} > 0$. For every partition $\mathcal{T}_0 = \{B_k \times C_\ell : k \in [K], \ell \in [L]\}$ of $[n] \times [m]$, let us define the projection matrix $\Pi_{\mathcal{T}_0}$ by*

$$(\Pi_{\mathcal{T}_0} \mathbf{V})_{i,j} = \frac{1}{|B_k||C_\ell|} \sum_{i' \in B_k} \sum_{j' \in C_\ell} V_{i',j'}\,; \qquad \textit{if} \quad (i,j) \in B_k \times C_\ell, \qquad \forall \mathbf{V} \in \mathbb{R}^{n \times m}.$$

1. *If $\|\mathbf{E}\|_\infty \leqslant 1$, then for all $t > 0$, we have that with probability at least $1 - e^{-t}$,*

$$\forall \mathbf{V},\ \langle \mathbf{V}, \Pi_{\mathcal{T}_0} \mathbf{E} \rangle \leqslant \frac{1}{2} \|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\mathsf{F}} \|\mathbf{V}\|_{\mathsf{F}} + \sqrt{2\rho_{\mathbf{\Sigma}}(t + \log M)} \|\mathbf{V}\|_{\mathsf{F}} + \frac{2(t+\log M)}{3} \|\mathbf{V}\|_{1,\infty},$$

   and on the same event, we have

$$\|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\mathsf{F}}^2 \leqslant 8(t + \log M)\big(\rho_{\mathbf{\Sigma}} + (1/3)\|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{1,\infty}\big).$$

2. *If $\|\mathbf{E}\|_{1,\infty} \leqslant \rho_{\mathbf{E}}$, then for all $t > 0$, we have that with probability at least $1 - e^{-t}$,*

$$\forall \mathbf{V},\ \langle \mathbf{V}, \Pi_{\mathcal{T}_0} \mathbf{E} \rangle \leqslant \frac{1}{2} \|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\mathsf{F}} \|\mathbf{V}\|_{\mathsf{F}} + \sqrt{2\rho_{\mathbf{\Sigma}}(t + \log M)} \|\mathbf{V}\|_{\mathsf{F}} + \frac{2(t+\log M)}{3} \rho_{\mathbf{E}} \|\Pi_{\mathcal{T}_0} \mathbf{V}\|_{\infty}$$

$$\textit{and} \quad \|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\mathsf{F}}^2 \leqslant 8(t + \log M)\big(\rho_{\mathbf{\Sigma}} + (1/3)\rho_{\mathbf{E}} \|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\infty}\big).$$

where $\log M \leqslant KL \log 12 \leqslant 2.5KL$.

Proof. We prove only the first part of the lemma, the second one is proved similarly. Let $\mathbf{W}$ be an $n \times m$ matrix of unit Frobenius norm. We have

$$\langle \mathbf{W}, \Pi_{\mathcal{T}_0} \mathbf{E} \rangle = \langle \Pi_{\mathcal{T}_0} \mathbf{W}, \mathbf{E} \rangle = \sum_{i=1}^{n} \boldsymbol{E}_i^\top (\Pi_{\mathcal{T}_0} \mathbf{W})_i.$$

For $i \in [n]$, the random variables $\boldsymbol{E}_i^\top (\Pi_{\mathcal{T}_0} \mathbf{W})_i$ are independent and satisfy $|\boldsymbol{E}_i^\top (\Pi_{\mathcal{T}_0} \mathbf{W})_i| \leqslant \|(\Pi_{\mathcal{T}_0} \mathbf{W})_i\|_1$ and

$$\sum_{i=1}^{n} \mathbb{E}[(\boldsymbol{E}_i^\top (\Pi_{\mathcal{T}_0} \mathbf{W})_i)^2] = \sum_{i=1}^{n} (\Pi_{\mathcal{T}_0} \mathbf{W})_i^\top \boldsymbol{\Sigma} (\Pi_{\mathcal{T}_0} \mathbf{W})_i \leq \rho_{\boldsymbol{\Sigma}} \sum_{i=1}^{n} \|(\Pi_{\mathcal{T}_0} \mathbf{W})_i\|_2^2 = \rho_{\boldsymbol{\Sigma}} \|\Pi_{\mathcal{T}_0} \mathbf{W}\|_{\mathsf{F}}^2 \leq \rho_{\boldsymbol{\Sigma}}.$$

The Bernstein inequality yields

$$\mathbb{P}\Big(\langle \mathbf{W}, \Pi_{\mathcal{T}_0}(\mathbf{E}) \rangle \leqslant \sqrt{2\rho_{\boldsymbol{\Sigma}} t} + \frac{2t}{3} \|\Pi_{\mathcal{T}_0} \mathbf{W}\|_{1,\infty} \Big) \geqslant 1 - e^{-t}, \qquad t \geqslant 0. \tag{3.15}$$

Let $V_{\mathcal{T}_0} = \{\Pi_{\mathcal{T}_0} \mathbf{A} : \|\mathbf{A}\|_{\mathsf{F}} = 1\}$, be the image of the unit ball of $\mathbb{R}^{n \times m}$ by $\Pi_{\mathcal{T}_0}$ and let $\mathcal{N}_{\mathcal{T}_0} = \{\mathbf{W}^1, \dots, \mathbf{W}^M\}$ be an ε-net of $V_{\mathcal{T}_0}$ for $\varepsilon = 1/4$. The set $V_{\mathcal{T}_0}$ being included in the unit ball of a linear space of dimension KL, [vH16, Lemma 5.13] shows that $M \leqslant 12^{KL}$. Define

$$\mathbf{U}^m = \operatorname*{arg\,min}_{\mathbf{U} \in V_{\mathcal{T}_0}} \Big\{ \|\mathbf{U}\|_{1,\infty} : \|\mathbf{U} - \mathbf{W}^m\|_{\mathsf{F}} \leq 1/4 \Big\}, \qquad m = 1, \dots, M.$$

Using (3.15) and the union bound, we get

$$\mathbb{P}\Big(\forall m \in [M],\ \langle \mathbf{U}^m, \Pi_{\mathcal{T}_0} \mathbf{E} \rangle \leqslant \sqrt{2\rho_{\boldsymbol{\Sigma}}(t + \log M)} + \frac{2(t + \log M)}{3} \|\mathbf{U}^m\|_{1,\infty} \Big) \geqslant 1 - e^{-t}, \quad t \geqslant 0,$$

where we used the fact that $\Pi_{\mathcal{T}_0} \mathbf{U}^m = \mathbf{U}^m$.

Let $\mathbf{W}$ be an arbitrary matrix from $V_{\mathcal{T}_0}$. Let $\mathbf{W}^m$ be any point from the net $\mathcal{N}_{\mathcal{T}_0}$ such that $\|\mathbf{W} - \mathbf{W}^m\|_2 \leq 1/4$. We have $\|\mathbf{U}^m\|_{1,\infty} \leq \|\mathbf{W}\|_{1,\infty}$ and $\|\mathbf{W} - \mathbf{U}^m\|_{\mathsf{F}} \leq 1/2$. This implies that, with probability at least $1 - e^{-t}$, for any matrix $\mathbf{W} \in V_{\mathcal{T}_0}$,

$$\begin{aligned}
\langle \Pi_{\mathcal{T}_0} \mathbf{W}, \mathbf{E} \rangle &= \langle \mathbf{W} - \mathbf{U}^m, \Pi_{\mathcal{T}_0} \mathbf{E} \rangle + \langle \mathbf{U}^m, \Pi_{\mathcal{T}_0} \mathbf{E} \rangle \\
&\leqslant \|\mathbf{W} - \mathbf{U}^m\|_{\mathsf{F}} \|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\mathsf{F}} + \langle \mathbf{U}^m, \Pi_{\mathcal{T}_0} \mathbf{E} \rangle \\
&\leqslant (1/2) \|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\mathsf{F}} + \sqrt{2\rho_{\boldsymbol{\Sigma}}(t + \log M)} + (2/3)(t + \log M) \|\mathbf{U}^m\|_{1,\infty} \\
&\leqslant (1/2) \|\Pi_{\mathcal{T}_0} \mathbf{E}\|_{\mathsf{F}} + \sqrt{2\rho_{\boldsymbol{\Sigma}}(t + \log M)} + (2/3)(t + \log M) \|\mathbf{W}\|_{1,\infty}.
\end{aligned}$$

Since this inequality is valid for any matrix $\mathbf{W}$ in the image of the unit ball by $\Pi_{\mathcal{T}_0}$, it is also valid for $\mathbf{W} = \Pi_{\mathcal{T}_0} \mathbf{V} / \|\Pi_{\mathcal{T}_0} \mathbf{V}\|_{\mathsf{F}}$. Replacing this with the last display, then multiplying the two sides of the inequality by $\|\Pi_{\mathcal{T}_0} \mathbf{V}\|_{\mathsf{F}}$, we get the first claim of the lemma.

For the second claim, we take $\mathbf{V} = \Pi_{\mathcal{T}_0}\mathbf{E}$ to get

$$\begin{aligned}\|\Pi_{\mathcal{T}_0}\mathbf{E}\|_2^2 &\leqslant 2\sqrt{2\rho_{\mathbf{\Sigma}}(t+\log M)}\|\Pi_{\mathcal{T}_0}\mathbf{E}\|_2 + (4/3)(t+\log M)\|\Pi_{\mathcal{T}_0}\mathbf{E}\|_{1,\infty}\\ &\leqslant (1/2)\|\Pi_{\mathcal{T}_0}\mathbf{E}\|_2^2 + 4\rho_{\mathbf{\Sigma}}(t+\log M) + (4/3)(t+\log M)\|\Pi_{\mathcal{T}_0}\mathbf{E}\|_{1,\infty},\end{aligned}$$

where we used the inequality $2uv \leqslant (1/2)u^2 + 2v^2$ for all $u, v \in \mathbb{R}$. Rearranging the terms of the last display, we obtain the claim of the lemma. □

Lemma 25. *Let $\boldsymbol{E}_1, \dots, \boldsymbol{E}_n$ be centered and independent random vectors in $\mathbb{R}^m$ such that $\|\boldsymbol{E}_i\|_1 \leqslant \rho_{\mathbf{E}}$ and $\|\mathbb{E}[\boldsymbol{E}_i\boldsymbol{E}_i^\top]\|_{\mathrm{op}} \leqslant \rho_{\mathbf{\Sigma}}$ with $\rho_{\mathbf{\Sigma}} > 0$, for every $i \in [n]$. For every $t > 0$, on an event of probability at least $1 - 2e^{-t}$, the inequality*

$$\Big|\sum_{i\in B}\frac{\boldsymbol{E}_i^\top \mathbf{1}_C}{|B||C|}\Big| \leqslant \sqrt{2\rho_{\mathbf{\Sigma}}\Big(\frac{t+\log(nm)}{n_0m_0} + \psi_{n,m}(n_0,m_0)\Big)} + \frac{2\rho_{\mathbf{E}}}{3}\Big(\frac{t+\log(nm)}{n_0m_0} + \psi_{n,m}(n_0,m_0)\Big)$$

holds true for every pair of subsets $B \subset [n]$ and $C \subset [m]$ such that $|B| \geq n_0$ and $|C| \geq m_0$.

Proof. The random variables $\boldsymbol{E}_i^\top \mathbf{1}_C$, $i \in B$ are independent, zero mean and satisfy

$$|\boldsymbol{E}_i^\top \mathbf{1}_C| \leq \rho_{\mathbf{E}}, \quad \mathbb{E}[(\boldsymbol{E}_i^\top \mathbf{1}_C)^2] = \mathbf{1}_C^\top \mathbf{\Sigma} \mathbf{1}_C \leq \rho_{\mathbf{\Sigma}}\|\mathbf{1}_C\|_2^2 = |C|\rho_{\mathbf{\Sigma}}.$$

In view of the Bernstein inequality, this yields

$$\mathbb{P}\Big(\Big|\sum_{i\in B}\boldsymbol{E}_i^\top \mathbf{1}_C\Big| \leqslant \sqrt{2|B||C|\rho_{\mathbf{\Sigma}}t} + \frac{2\rho_{\mathbf{E}}t}{3}\Big) \geqslant 1 - 2e^{-t}, \qquad \forall t > 0.$$

This implies that

$$\mathbb{P}\Big(\Big|\sum_{i\in B}\frac{1}{|B||C|}\boldsymbol{E}_i^\top \mathbf{1}_C\Big| \leqslant \sqrt{\frac{2\rho_{\mathbf{\Sigma}}t}{|B||C|}} + \frac{2\rho_{\mathbf{E}}t}{3|B||C|}\Big) \geqslant 1 - 2e^{-t}, \qquad \forall t > 0.$$

Let $n_B \in [n]$ and $m_C \in [m]$ be two integers. The union bound combined with the last display leads to

$$\mathbb{P}\Big(\Big|\sum_{i\in B}\frac{\boldsymbol{E}_i^\top \mathbf{1}_C}{|B||C|}\Big| \leqslant \sqrt{\frac{2\rho_{\mathbf{\Sigma}}t}{|B||C|}} + \frac{2\rho_{\mathbf{E}}t}{3|B||C|}\Big) \geqslant 1 - 2\binom{n}{n_B}\binom{m}{m_C}e^{-t}.$$

Replacing t by

$$t + \log\binom{n}{n_B} + \log\binom{m}{m_C} \leqslant t + n_B\log\big(ne/n_B\big) + m_C\log\big(me/m_C\big)$$

we obtain that with probability at least $1 - 2e^{-t}$,

$$\left| \sum_{i \in B} \frac{\boldsymbol{E}_i^\top \mathbf{1}_C}{|B||C|} \right| \leqslant \sqrt{2\rho_{\boldsymbol{\Sigma}} \left(\frac{t}{n_B m_C} + \psi_{n,m}(n_B, m_B) \right)} + (2/3)\rho_{\mathbf{E}} \left(\frac{t}{n_B m_C} + \psi_{n,m}(n_B, m_B) \right)$$

for all B and C such that $|B| = n_B$ and $|C| = m_C$. Applying once again a union bound over all integers $n_B \in [n_0, n]$ and $m_C \in [m_0, m]$, we check that the inequality

$$\left| \sum_{i \in B} \frac{\boldsymbol{E}_i^\top \mathbf{1}_C}{|B||C|} \right| \leqslant \sqrt{2\rho_{\boldsymbol{\Sigma}} \left(\frac{t + \log(nm)}{n_0 m_0} + \psi_{n,m}(n_0, m_0) \right)} + (2/3)\rho_{\mathbf{E}} \left(\frac{t + \log(nm)}{n_0 m_0} + \psi_{n,m}(n_0, m_0) \right)$$

holds true for every pair of sets B and C such that $|B| \geq n_0$ and $|C| \geq m_0$ with probability at least $1 - 2e^{-t}$. □

Chapter 4

Conclusion

Summary In this thesis, we investigated the graphon estimation problem through the observation of a bipartite graph generated from this graphon and some unobserved latent variables, under two main assumptions.

We started by supposing that all the edges of the observed graph are labeled independently, that is all the entries of the observed adjacency matrix are independent, and drawn from some distribution that satisfies Bernstein conditions. We provided finite sample risk upper-bounds for both mean matrix and graphon least square estimators. Two main classes of graphon regularity has been considered, the piecewise constant graphons, which generate a bipartite stochastic block model, and the Hölder graphons. For the class of piecewise constant graphon, we proved that the least square estimator is optimal in the sens that the lower bound on the minimax risk is of the order of the upper-bound for many settings, when the entries of the adjacency matrix are drawn from general Binomial distributions knowing the latent variables.

In a second step, we proposed to relax the independence assumption about edges labeling, in order to model some situations with controlled number of edges for instance. In this framework, the labeled are assumed to be in the segment $[0, 1]$. We also provided upper-bounds for the mean matrix and the graphon estimations, only for the class of piecewise constant graphons. We did not neither investigate the case of Hölder graphons nor the optimality of our upper-bounds.

Matching lower and upper bounds for piecewise constant graphons Our research opens up several avenues for further investigation and raises intriguing questions. One important question relates to the optimality of the upper and lower bounds derived in our study for the estimation of piecewise constant graphons. Although these bounds are optimal for various parameter settings, there is a noticeable discrepancy between them, particularly due to the appearance of logarithmic factors in the upper bound. This gap persists even in the symmetric setting, and it is not specific to the bipartite framework we considered. This sub-optimality has been observed in previous research, such as [KTV17], which focuses on the unipartite setting.

Additionally, in asymmetric settings, the lower bound fails to match the upper bound due to the presence of the term $\sqrt{\frac{1}{nm}}$ in the lower bound, which can be significantly smaller than the term $\frac{1}{n} + \frac{1}{m}$ in the upper bound. For instance, when considering a scenario with $m = Cn$ with

C a positive constant, potentially large, the upper bound is approximately $\sqrt{C}$-times greater than the lower bound.

Lower bound for general α-Hölder continuous graphons The question of deriving lower bounds for Hölder graphon estimation is indeed an important and interesting one. As discuss in Theorem 8, the upper bound is known to be optimal (up to a $\log n$ factor) for Lipschitz continuous graphons, that is when $\alpha = 1$. In the case of Hölder continuous graphons, the upper bound we derived in Theorem 8 is of the order $n^{-\alpha/2}$, but it is not known whether this upper bound represents the optimal rate for estimation. To establish the optimality of the least squares estimator for Hölder graphons, it would be necessary to derive corresponding lower bounds that match or closely approach this upper bound.

Hölder graphon estimation in the relaxed independence setting In Chapter 3, we focused on deriving upper bounds for estimating piecewise constant graphons under a relaxed independence assumption. However, extending these bounds to Hölder continuous graphons poses additional challenges due to identifiability issues and the presence of normalization in the distance definition.

The main issue arises from the fact that our model does not distinguish between a graphon W and its normalized version $I_W^{-1}W$, where $I_W(x) = \int_0^1 W(x,y)dy$. In the case of general graphons, I_W can be very small, leading to a large approximation error that is reflected in the upper bound. This makes it difficult to adapt the proofs from the full independence framework to Hölder continuous graphons.

More general distributions for the adjacency matrix (relaxed independence setting) In the relaxed independence setting, considering more general distributions of adjacency matrices, including unbounded distributions, could indeed be a promising direction. One possible approach is to explore (a,b)-Bernstein conditions, building upon the ideas discussed in Chapter 2. These conditions could potentially address the sub-optimality issues related to noise and sparsity parameters in the upper bound. However, for technical reasons, we did not achieve to obtain satisfying upper bounds in this framework.

Explore other class of regularity Considering monotonicity assumptions for graphons can be a valuable direction for future research. By imposing monotonicity constraints, such as the assumption that the graphon is increasing in its first argument, we can certainly enhance identifiability.

Assuming that the graphon is increasing in its first argument, that is $x \mapsto W(x,y)$ is increasing, implies that as the unobserved characteristic U_i of the i-th node increases, the probability for that node to form links also increases. This assumption aligns with realistic scenarios where certain attributes or characteristics of nodes influence their connectivity patterns. From a mathematical standpoint, we did not delve into the exploration of this framework.

We can also study the cases of α-Hölder regularity with $\alpha > 1$.

Chapter 5

Résumé en français

5.1 Définition du problème

Considérations préliminaires Dans le domaine de l'économie, les ensembles de données de réseaux sont fréquemment utilisés pour modéliser les marchés ou les interactions entre différentes entités ou individus. En conséquence, il existe une vaste littérature sur le sujet des réseaux économiques et des modèles de formation de liens, comme en témoignent les travaux de [Gra17, Gra20, DG14b, JW96, Dze19]. Les réseaux unipartites et bipartites peuvent être pertinents pour la modélisation économique. Par exemple, les échanges commerciaux internationaux peuvent être modélisés à l'aide d'un graphe unipartite où les liens représentent l'existence d'échanges commerciaux. En revanche, les graphes bipartites sont plus appropriés pour modéliser les achats de produits par les consommateurs ou l'embauche de travailleurs dans une entreprise.

Lors de la modélisation de la formation de réseaux, il y a généralement deux types de variables à prendre en compte : les variables observables et les variables non observables.

- Les variables observables sont des caractéristiques auxquelles nous pouvons accéder ou que nous pouvons calculer à partir des données que nous observons. Par exemple, dans le contexte du commerce international, nous pouvons avoir des informations sur la taille, le PIB ou la localisation géographique de chaque pays.
- En revanche, les variables non observables sont des variables latentes que nous ne pouvons pas observer ou calculer directement. Dans le cas du réseau de travailleurs et d'entreprises, des exemples de variables non observables comprennent par exemple la sympathie d'un travailleur, qui peut influencer le processus d'embauche, ou l'attractivité (la bonne ambiance de travail) d'une entreprise, qui peut influencer les choix des travailleurs.

[Gra17] propose un modèle économétrique de formation de réseau pour un graphe unipartite avec à la fois des variables observées et non observées. En revanche, [Gra20] présente un modèle logistique pour un réseau bipartite avec uniquement des variables observées.

Maintenant se pose la question des hypothèses concernant la formation des liens. L'hypothèse la plus courante est que tous les liens sont formés de manière indépendante, ce qui est plus ou moins réaliste ou parfois même non pertinent en fonction de la situation de modélisation, mais

cela simplifie les problèmes mathématiques à traiter. Une hypothèse plus réaliste et plus faible est décrite dans [DDG21a], qui parle d'échangeabilité. Grossièrement parlant, cela signifie que nous pouvons permuter les étiquettes des sommets sans changer la distribution des arêtes, ou de manière équivalente, nous pouvons appliquer n'importe quelle permutation des indices de lignes et de colonnes à la matrice d'adjacence de notre graphe sans changer sa loi.

Selon [Ald81, Théorème 1.4], un graphe biparti avec une matrice d'adjacence échangeable en lignes et en colonnes $\mathbf{A}$ peut être représenté par une fonction $g^* : [0,1]^4 \to \mathbb{R}$ et des variables aléatoires indépendantes et uniformément distribuées α, U_i, V_j, ξ_{ij} dans $[0,1]$, de la manière suivante :

$$\mathbf{A} \overset{(\mathcal{L})}{=} \big(g^*(\alpha, U_i, V_j, \xi_{ij}); i \in [n], j \in [m]\big)$$

Ici, les variables aléatoires U_i, V_j et $\xi_{i,j}$ correspondent aux variables non observables mentionnées précédemment. Si des vecteurs de caractéristiques observables supplémentaires $\boldsymbol{X}_{i,j}$ sont disponibles pour chaque paire de nœuds (i,j), alors un modèle étendu peut être envisagé, défini par $g^*(\boldsymbol{X}_{i,j}, \alpha, U_i, V_j, \xi_{i,j})$. Notre objectif est d'estimer la fonction g^* à partir du graphe qu'elle génère, mais cette tâche peut être complexe. Pour simplifier le problème, nous proposons de ne considérer que les variables non observables en supprimant toutes les variables observables. Dans une première étape, nous supposons que les arêtes sont formées de manière indépendante conditionnellement aux variables non observables. Nous définissons formellement le problème ci-dessous. Dans une deuxième étape, nous souhaitons assouplir l'hypothèse d'indépendance. En effet, si nous considérons le problème du réseau travailleur-entreprise, un travailleur doit choisir une et une seule entreprise pour travailler, ce qui signifie que les liens ne sont plus indépendants.

Définition mathématique du problème principal Soient n et m deux entiers positifs supposés grands, et $\mathbf{H}$ une matrice aléatoire de dimensions $n \times m$ avec des entrées réelles $H_{i,j}$. La matrice $\mathbf{H}$ peut être vue comme la matrice d'adjacence d'un graphe biparti avec des étiquettes d'arêtes. Nous supposons que la distribution de cette matrice $\mathbf{H}$ satisfait la condition suivante.

Hypothèse 1 (Indépendance totale)**.** Il existe une fonction $W^* : [0,1]^2 \to \mathbb{R}$, appelée le graphon, et deux vecteurs aléatoires $\boldsymbol{U} = (U_1, \dots, U_n)$ et $\boldsymbol{V} = (V_1, \dots, V_m)$ tels que

H 1.1 Les variables $U_1, \dots, U_n, V_1, \dots, V_m$ sont indépendantes et suivent une distribution uniforme $\mathcal{U}([0,1])$.

H 1.2 Conditionnellement à $(\boldsymbol{U}, \boldsymbol{V})$, les entrées $H_{i,j}$ sont indépendantes et $\mathbb{E}[H_{i,j}|\boldsymbol{U}, \boldsymbol{V}] = W^*(U_i, V_j)$.

Cette hypothèse 1 doit être comprise comme suit. Chaque sommet du côté gauche du graphe biparti est associé à une variable non observée U_i, et de même pour le côté droit avec les variables V_j. De plus, si nous connaissions les variables U_i et V_j, nous supposons que les entrées de la matrice d'adjacence $\mathbf{H}$ sont indépendantes et suivent la loi de Bernoulli de paramètre $W^*(U_i, V_j)$, ce qui signifie qu'une arête entre i et j est présente avec une probabilité $W^*(U_i, V_j)$. Des distributions plus générales seront considérées ultérieurement, permettant des étiquettes d'arêtes.

Nous avons pour objectif d'étudier le risque minimax de l'estimation du graphon W^* à partir de l'observation de $\mathbf{H}$, et de démontrer comment il dépend de paramètres cruciaux du problème. Bien que les dimensions de la matrice n et m figurent parmi ces paramètres, nous explorons également l'impact de la régularité de W^*, du degré de "parcimonie" ou "sparsité" des interactions (représenté par ρ) et du niveau de bruit (représenté par σ). Pour être plus précis, σ et ρ sont des nombres réels positifs tels que

$$\|W^*\|_\infty = \sup_{u,v\in[0,1]} |W^*(u,v)| \leqslant \rho$$
$$\text{et} \qquad \mathbf{Var}[H_{i,j}|U_i, V_j] \leqslant \sigma^2 \text{ a.s.}, \quad \forall i \in [n], \forall j \in [m].$$

Pour fixer les idées, supposons que W^* est en escalier (constant sur certains rectangles du carré unité $[0,1]^2$) et définissons $\mathbf{\Theta}^* \in [0,1]^{n\times m}$ la matrice aléatoire

$$\Theta^*_{i,j} = W^*(U_i, V_j).$$

Remarquez que $\mathbf{\Theta}^*$ est constant par blocs, à quelques permutations près des lignes et des colonnes. Supposons de plus que conditionnellement à $(\boldsymbol{U}, \boldsymbol{V})$, les entrées $H_{i,j}$ sont indépendamment tirées de la loi de Bernoulli de paramètre $\Theta^*_{i,j}$. Ce scénario correspond stochastic block model présenté dans l'introduction, où la matrice de probabilité associée $\mathbf{\Theta}^*$ est inconnue. Ainsi, le problème d'estimation de W^* est équivalent au problème de détection de communauté, où nous cherchons à estimer les clusters et la matrice de probabilité. En plus de la distribution de Bernoulli, nous visons à fournir une borne supérieure du risque pour notre méthode d'estimation pour des distributions plus générales. Dans cette thèse, nous examinerons également un cadre non paramétrique : la classe des graphons réguliers α-Hölder comme une autre forme de régularité des graphons. Pour ce faire, nous visons à approximer les graphons Hölder par des graphons en escalier (voir figure 21).

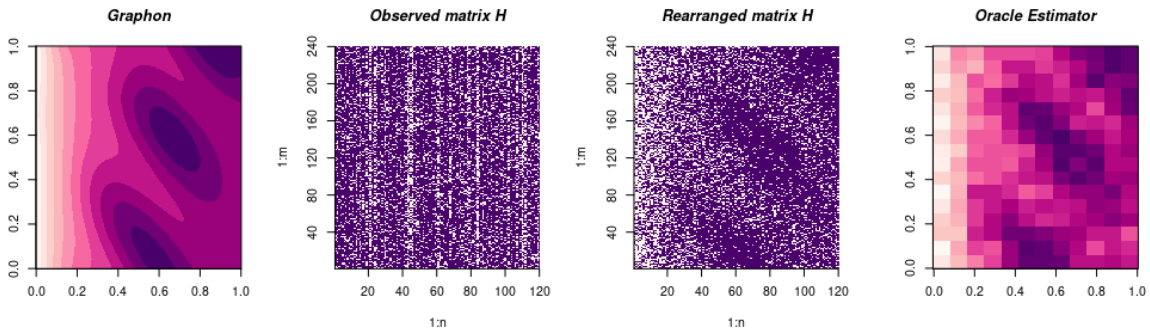


Figure 21: Une illustration du problème de graphon. La figure la plus à gauche représente le graphon inconnu W^*. La deuxième figure la plus à gauche est la matrice d'adjacence observée dans le graphe où les liens sont formés selon le modèle de Bernoulli. La troisième figure est la matrice d'adjacence qui serait obtenue après un réarrangement des lignes et des colonnes si nous avions accès aux variables latentes. La figure la plus à droite représente l'estimateur par histogramme obtenu à partir de la matrice d'adjacence réarrangée. Notre objectif est de concevoir un estimateur qui soit presque aussi performant que l'oracle, sans avoir accès aux variables latentes.

Relâchement de l'hypothèse d'indépendance Le cadre précédent, qui suppose l'indépendance entre les arêtes, peut ne pas convenir pour modéliser certaines situations courantes rencontrées dans la pratique. Un exemple de ce type est le réseau travailleur-entreprise, où le premier ensemble de sommets représente les travailleurs et le deuxième ensemble représente les entreprises. Un travailleur est connecté à une entreprise s'il est embauché par cette dernière. Dans ce scénario, il est raisonnable de supposer que chaque travailleur est embauché par au plus une entreprise, ce qui entraîne un degré maximal de 1 pour chaque sommet dans le premier ensemble. L'idée est alors de relâcher l'hypothèse d'indépendance concernant la formation des liens. Nous ne considérons également que des variables latentes non observées dans ce modèle, et supposons que la matrice d'adjacence $\mathbf{H}$ vit maintenant dans $[0,1]^{n\times m}$ et satisfait l'énoncé suivant.

Hypothèse 2 (Indépendance relâchée)**.** Nous considérons une fonction $W^* : [0,1]^2 \to [0,+\infty[$ appelée le graphon, et deux vecteurs aléatoires $\boldsymbol{U} = (U_1,\dots,U_n)$ et $\boldsymbol{V} = (V_1,\dots,V_m)$ qui satisfont

H 2.1 $U_1,\dots,U_n,V_1,\dots,V_m$ sont indépendants et suivent une distribution uniforme sur $[0,1]$.

H 2.2 Conditionnellement à $(\boldsymbol{U},\boldsymbol{V})$, les lignes de la matrice $\mathbf{H}$ sont indépendantes.

H 2.3 Chaque ligne de $\mathbf{H}$ somme à un et

$$\mathbb{E}[H_{i,j}|\boldsymbol{U},\boldsymbol{V}] = \frac{W^*(U_i,V_j)}{\sum_{\ell=1}^m W^*(U_i,V_\ell)}.$$

L'hypothèse **H 2.1** est la même que dans la hypothèse 1, où nous considérons que les variables non observées attribuées à chaque sommet sont indépendantes. Cependant, l'hypothèse **H 2.2** relâche l'hypothèse d'indépendance concernant les liens formés par les individus du côté droit. Au lieu de cela, elle exige uniquement que les lignes de $\mathbf{H}$ soient conditionnellement indépendantes étant donné $(\boldsymbol{U},\boldsymbol{V})$. En d'autres termes, alors que les arêtes formées par des individus distincts du côté droit sont indépendantes, les liens formés par un seul individu ne sont pas nécessairement indépendants.

Notre objectif est double. Tout d'abord, nous visons à estimer la matrice moyenne de $\mathbf{H}$ et à fournir une borne de risque pour notre méthode d'estimation. Dans cette partie, nous ne considérons pas la partie **H 2.3** de l'hypothèse hypothèse 2 qui pourrait être restrictive, mais nous la remplaçons plutôt par l'hypothèse suivante :

$$\sum_{j=1}^m H_{i,j} \leqslant \rho_{\mathbf{H}} \quad \forall i = 1,\dots,m \qquad \textbf{(A 2.3} \text{ (bis))}$$

c'est-à-dire, la somme de chaque ligne de $\mathbf{H}$ est inférieure à un paramètre positif $\rho_{\mathbf{H}}$. Nous considérerons souvent $\rho_{\mathbf{H}} = 1$. Notre objectif est d'analyser comment le risque de notre estimateur se comporte par rapport aux paramètres du modèle, en particulier n et m, la taille de notre ensemble de données, ainsi que les paramètres $\rho_{\mathbf{H}}$, qui représentent la contrainte de somme des lignes, et $\rho_{\boldsymbol{\Sigma}}$, le niveau de bruit, qui satisfait maintenant[1]

$$\|\boldsymbol{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\boldsymbol{\Sigma}} \quad \forall i = 1,\dots,n \tag{5.1}$$

[1] $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$ fait référence à la norme d'opérateur d'une matrice carrée.

où $\mathbf{\Sigma}_i = \mathbb{E}[\boldsymbol{H}_i \boldsymbol{H}_i^\top] - \mathbb{E}[\boldsymbol{H}_i]\mathbb{E}[\boldsymbol{H}_i]^\top$ est la matrice de covariance de la i-ème ligne de $\mathbf{H}$.

Il sera courant de supposer que la matrice $\mathbf{\Theta}^* = \mathbb{E}[\mathbf{H}]$ a des lignes dont la somme est bornée par $\rho_{\mathbf{\Theta}} > 0$. Ce paramètre peut également apparaître dans les bornes supérieures du risque. La deuxième partie de notre travail tourne autour de l'estimation du graphon W^* sous l'hypothèse hypothèse 2 et la condition (5.1). De plus, nous nous concentrons exclusivement sur la classe des graphons en escalier pour cette tâche d'estimation.

5.2 Contributions

Nous présentons un résumé des principales contributions de cette thèse en quatre sous-sections.

- La première sous-section se concentre sur le problème d'estimation de la moyenne, qui est crucial dans les deux contextes d'indépendance décrits dans les hypothèses 1 and 2. Cette étape est significative et intéressante en soi.
- La deuxième sous-section aborde le problème d'estimation du graphon, en s'appuyant sur la procédure d'estimation dérivée de la première étape.
- Dans la troisième sous-section, nous établissons des bornes inférieures sur le risque dans le pire des cas pour tout estimateur de graphon dans l'ensemble des graphons en escalier, sous l'hypothèse de liens complètement indépendants. Ces bornes inférieures s'appliquent lorsque les matrices ont des entrées avec une loi conditionnelle binomiale étant donné les variables latentes. Remarquablement, dans la plupart des cas, ces bornes inférieures sont du même ordre que les bornes supérieures obtenues pour l'estimateur des moindres carrés.
- Enfin, dans la quatrième sous-section, nous présentons une adaptation de l'algorithme de minimisation alternative de Lloyd, incorporant une étape de relaxation convexe, à notre contexte spécifique. Cette adaptation nous permet d'obtenir une approximation calculable de l'estimateur des moindres carrés, et certaines simulations basées sur des données synthétiques, uniquement pour le contexte d'indépendance complète.

Pour simplifier les énoncés de théorèmes, nous adopterons souvent le cadre symétrique, où les deux côtés des graphes bipartites ont un nombre égal de sommets, de communautés, etc.

5.2.1 Estimation de la matrice moyenne

Hypothèse d'indépendance complète Sur la voie de l'estimation du graphon W^*, une étape intermédiaire importante consistera à estimer la matrice $\mathbf{\Theta}^* = W^*(U_i, V_j)$. L'estimation de cette matrice est intéressante en elle-même. Nous accomplissons cette tâche en résolvant le problème des moindres carrés sur l'ensemble des matrices constantes par blocs, avec des blocs générés par des partitions des ensembles de lignes et des colonnes de la matrice $\mathbf{H}$. Il sera également démontré que la méthode d'agrégation par des poids exponentiels peut être utilisée pour assurer l'adaptabilité au nombre de blocs. Sous la condition que le graphon soit

en escalier ou α-régulier au sens de la régularité de Hölder, nous établissons des bornes de risque pour l'estimateur de graphon dérivé de l'estimateur de $\mathbf{\Theta}^*$. Ces bornes de risque sont non asymptotiques et se révèlent être optimales au sens minimax pour une large gamme de régimes.

Les estimateurs des moindres carrés de $\mathbf{\Theta}^*$ sont définis comme la meilleure approximation de $\mathbf{H}$ par une matrice constante par blocs. Pour être plus précis,

$$\widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K,L] \in \arg\min_{\mathbf{\Theta}\in\mathcal{T}^{K,L}_{n_0,m_0}} |\mathbf{H}-\mathbf{\Theta}|\mathsf{F}^2. \tag{5.2}$$

Ici, $\mathcal{T}n_0, m_0{}^{K,L}$ représente un ensemble de matrices constantes par blocs avec $K \times L$ blocs. Les paramètres $n_0 \geqslant 1$ et $m_0 \geqslant 1$ font référence au nombre minimal d'entrées dans chaque bloc. Nous dérivons des bornes de risque pour les estimateurs de $\mathbf{\Theta}^*$. Dans le cadre de notre analyse, nous considérons $\mathbf{\Theta}^*$ comme une matrice déterministe, ce qui nous permet de supposer l'indépendance de $H_{i,j}$ plutôt que l'indépendance conditionnelle étant donné $\boldsymbol{U}$ et $\boldsymbol{V}$. De plus, dans un but de simplicité, dans l'énoncé du prochain théorème, nous faisons l'hypothèse de symétrie, où $n = m$, $K = L$ et $n_0 = m_0$.

Théorème 1. *Soient n, n_0 et K des entiers positifs tels que $K \geqslant 2$ et $3 \leqslant n_0 \leqslant n$. Soit $\mathbf{H}$ une matrice aléatoire de taille $n \times n$ avec des entrées indépendantes satisfaisant $\mathbb{E}[H_{ij}] \in [0,\rho]$ pour chaque $i,j \in [n]$ et pour un certain $\rho > 0$. De plus, supposons que les variables aléatoires $(H_{ij} - \mathbb{E}[H_{ij}])$ satisfont la condition (σ^2, b)-Bernstein [2]. Alors, l'estimateur des moindres carrés $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ de la matrice moyenne $\mathbf{\Theta}^* = \mathbb{E}[\mathbf{H}]$, défini par (5.2), satisfait l'inégalité oracle exacte*

$$\frac{1}{n}\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2} \leqslant \inf_{\mathbf{\Theta}\in\mathcal{T}^K_{n_0}} \frac{1}{n}\|\mathbf{\Theta}-\mathbf{\Theta}^*\|\mathsf{F} + \big(25\sigma^2 + 4b\rho\big)^{1/2}\Big(\frac{3K^2}{n^2} + \frac{2\log K}{n}\Big)^{1/2},$$

à condition que $\psi_n(n_0) := \frac{6}{n_0}, \log(en/n_0) \leqslant (\sigma/b)^2$.

Dans la table 5.1, nous fournissons quatre exemples principaux illustrant les conséquences du théorème 1 dans le cas non symétrique pour des distributions courantes de $H_{i,j}$.

Il convient de noter que dans le cas symétrique, nous retrouvons la borne supérieure obtenue pour les graphes unipartites telle que décrite dans [KTV17]. Cela indique que les résultats obtenus dans le contexte actuel étendent et sont en accord avec les conclusions dans le cadre des graphes unipartites.

L'estimateur des moindres carrés $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ présente un motif constant par blocs avec KL blocs. Le choix de KL comme nombre de blocs est un hyperparamètre de la méthode. Cependant, si la vraie matrice $\mathbf{\Theta}^*$ s'éloigne considérablement d'une forme constante par blocs avec KL blocs, la qualité de l'estimation de $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0,m_0}[K,L]$ peut se détériorer en raison d'un biais important. Pour atténuer ce biais, une approche consiste à calculer l'estimateur des moindres carrés pour plusieurs valeurs de K, L, n_0 et m_0, puis à agréger ces estimateurs. En procédant ainsi, le biais peut être réduit et les performances globales de l'estimation peuvent être améliorées. Nous fournissons également une borne de risque à échantillon fini pour ce type d'estimateur agrégé. Enfin, les résultats mathématiques peuvent être aisément adaptés au cas des observations

[2] Une variable aléatoire centrée ζ satisfait la condition de Bernstein de paramètres (a,b) si $\mathbb{E}[e^{\lambda\zeta}] \leqslant \exp\big\{\frac{\lambda^2 a}{2(1-b|\lambda|)}\big\}$ tant que $|\lambda| \leqslant 1/b$.

Modèle	Définition	(σ^2, b)	Bornes supérieures
Bernoulli	$H_{i,j} \sim \mathcal{B}(\Theta^{i,j})$	$(\rho, 1/3)$	$9\sqrt{\rho}, \Big(\frac{KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2}$
Binomial	$NHi,j \sim \mathcal{B}(N, \Theta^{i,j})$	$(\rho/N, 1/3N)$	$9\sqrt{\rho}, \Big(\frac{KL}{Nnm} + \frac{\log K}{Nm} + \frac{\log L}{Nn}\Big)^{1/2}$
Poisson	$THi,j \sim \mathcal{P}(T\Theta)_{i,j}$	$(\rho/T, 1/3T)$	$9\sqrt{\rho}, \Big(\frac{KL}{Tnm} + \frac{\log K}{Tm} + \frac{\log L}{Tn}\Big)^{1/2}$
Sous-Gaussien	$\mathbb{E}[e^{\lambda H_{i,j}}] \leqslant e^{\sigma^2\lambda^2}$	$(\sigma^2, 0)$	$5\sigma, \Big(\frac{KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n}\Big)^{1/2}$

Table 5.1: Voici un résumé du deuxième terme obtenu pour les bornes supérieures dans la version non symétrique du théorème 1 pour des exemples spécifiques de distributions (σ^2, b)-Bernstein. Dans tous les cas, nous supposons que $\Theta^*_{i,j} \leqslant \rho$, sauf pour le modèle sous-gaussien.

manquantes, où certaines valeurs de la matrice $H_{i,j}$ ne sont pas observées.

Hypothèse de l'indépendance relachée Dans le cas de l'hypothèse de l'indépendance relachée, nous sommes en mesure de déduire une borne supérieure comparable. Cependant, il est important de noter que les distributions appliquées dans Théorème 2 ne sont pas aussi générales que celles supposées pour une indépendance totale dans le cadre précédent. Bien que la portée puisse être plus étroite, ces distributions fournissent néanmoins des informations et des résultats précieux pour notre analyse. Encore une fois, nous énonçons le théorème suivant uniquement dans le cadre symétrique.

Théorème 2. *Soient n, n_0 et K des entiers positifs tels que $K \geqslant 2$ et $1 \leqslant n_0 \leqslant n$. Soit $\mathbf{H} \in [0,1]^{n\times n}$ une matrice aléatoire de taille $n \times n$ avec des lignes indépendantes telles que la somme de chaque ligne soit égale à un et ayant une matrice de covariance $\mathbf{\Sigma}_i$ satisfaisant $\|\mathbf{\Sigma}_i\|_{\mathrm{op}} \leqslant \rho_{\mathbf{\Sigma}}$. Nous supposons également que $\|\mathbf{\Theta}^*\|_\infty \leqslant \rho_\infty$.*

L'estimateur des moindres carrés $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$ défini par (5.2) satisfait l'inégalité oracle exacte suivante :

$$\frac{1}{n}\mathbb{E}\big[\|\widehat{\mathbf{\Theta}}^{\mathsf{LS}} - \mathbf{\Theta}^*\|_{\mathsf{F}}^2\big]^{1/2} \leqslant \inf_{\mathbf{\Theta}\in\mathcal{T}} \frac{1}{n}\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_{\mathsf{F}} + \big(48\rho_{\mathbf{\Sigma}} + 6\rho_\infty\big)^{1/2}\Big(\frac{3K^2}{n^2} + \frac{2\log K}{n}\Big)^{1/2}$$

à condition que $\psi_n(n_0) = \frac{2\log(ne/n_0)}{n_0} \leqslant \rho_{\mathbf{\Sigma}}$.

Ce résultat englobe le cas précédemment décrit où le vecteur $\boldsymbol{H}_i$ a une seule entrée égale à 1, et les autres sont nulles, ce qui modélise une mise en correspondance où les individus du côté gauche doivent choisir un et un seul élément du côté droit, comme dans le réseau travailleur-entreprise.

5.2.2 Estimation du graphon

Hypothèse d'indépendance totale L'illustration dans la figure 21 met en évidence l'impact de la connaissance manquante des variables latentes sur le problème d'estimation du graphon.

Elle montre que la matrice d'adjacence réarrangée, obtenue si les variables latentes étaient connues, fournit des informations significativement plus précises sur le véritable graphon W^* par rapport à la matrice d'adjacence originale. Lorsque les variables latentes $\boldsymbol{U}$ et $\boldsymbol{V}$ sont inconnues, le graphon W^* devient non identifiable. Nous définissons l'équivalence entre deux graphons W et W' s'il existe deux bijections $\tau_1 : [0,1] \to [0,1]$ et $\tau_2 : [0,1] \to [0,1]$ qui préservent la mesure de Lebesgue, telles que $W = W' \circ (\tau_1 \otimes \tau_2)$ [3]. On peut observer que deux matrices $\mathbf{H}$ générées par des graphons équivalents W^* et $\bar{W}^*$ ont la même distribution. Par conséquent, la meilleure chose que nous puissions faire est d'estimer la classe d'équivalence contenant W^*. Cela motive l'utilisation de la (pseudo) distance employée dans ce travail pour évaluer la qualité d'un estimateur $\widehat{W}$ de W^*, comme suit :

$$\begin{aligned}\delta(\widehat{W}, W^*) &= \inf_{\tau_1,\tau_2 \in \mathcal{M}} \Big(\iint_{[0,1]^2} \big|\widehat{W}(\tau_1(u), \tau_2(v)) - W^*(u,v)\big|^2 du\, dv \Big)^{1/2} \\ &= \inf_{\tau_1,\tau_2 \in \mathcal{M}} \|\widehat{W} \circ (\tau_1 \otimes \tau_2) - W^*\|_{\mathbb{L}^2}\end{aligned}$$

où $\mathcal{M}$ est l'ensemble de toutes les automorphismes $\tau : [0,1] \to [0,1]$ tels que τ et τ^{-1} sont mesurables, et τ préserve la mesure de Lebesgue au sens où $\lambda(\tau^{-1}(B)) = \lambda(B)$ pour chaque boréliens $B \subset [0,1]$.

Après avoir estimé la matrice $\boldsymbol{\Theta}^*$ et choisi une mesure de distance pour l'évaluation de la qualité du graphon, l'étape suivante consiste à concevoir un estimateur pour le graphon W^*. Pour ce faire, nous associons un graphon $W_{\boldsymbol{\Theta}}$ à toute matrice $\boldsymbol{\Theta}$ de taille $n \times m$, où $W_{\boldsymbol{\Theta}} : [0,1]^2 \to [0,1]$ est défini comme une fonction constante sur chaque rectangle $I_i \times J_j = \left[\frac{i-1}{n}, \frac{i}{n}\right) \times \left[\frac{j-1}{m}, \frac{j}{m}\right)$ pour $(i,j) \in [n] \times [m]$:

$$W_{\boldsymbol{\Theta}}(u,v) = \Theta_{i,j}, \quad \text{pour tout} \quad (u,v) \in I_i \times J_j.$$

Dans le théorème qui suit, nous analyserons l'estimateur $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}}$. Comme mentionné précédemment, nous classerons W^* en deux catégories en fonction de sa régularité : la classe des graphons constant par morceaux et la classe $\mathbb{H}_{\alpha,\mathcal{L}}$ [4]. L'énoncé présenté ici traite spécifiquement du cas symétrique simplifié. Cependant, il est important de noter que nous avons également obtenu des résultats pour le cas asymétrique, qui sont discutés en détail dans le Chapter 2. Un résumé de ces résultats peut être trouvé dans table 5.2.

Théorème 3. *Soit $\mathbf{H}$ une matrice aléatoire de taille $n \times n$ satisfaisant l'Hypothèse 1 avec un certain graphon $W^* : [0,1]^2 \to [0,\rho]$. Supposons que pour une constante $\sigma > 0$, conditionnellement à $\boldsymbol{U}, \boldsymbol{V}$, les variables aléatoires $(H_{ij} - \mathbb{E}[H_{ij}|\boldsymbol{U},\boldsymbol{V}])$ satisfont la condition (σ^2, b)-Bernstein.*

1. *Supposons que le graphon W^* est K-piecewise constant, c'est-à-dire que pour un certain entier $K \geqslant 2$ et pour $0 = a_0 < \ldots < a_K = 1$ tel que*

$$\Delta^{(K)} := \min_{k \in [K]} |a_k - a_{k-1}| \geqslant \frac{8 \log(nK)}{n}$$

[3] Nous utilisons la notation $\tau_1 \otimes \tau_2$ pour la fonction de $[0,1]^2$ à $[0,1]^2$ définie par $(\tau_1 \otimes \tau_2)(u,v) = (\tau_1(u), \tau_2(v))$.

[4] $\mathbb{H}_{\alpha,\mathcal{L}}$ est l'ensemble des fonctions $W : [0,1]^2 \to \mathbb{R}$ satisfaisant $|W(x,y) - W(x',y')| \leqslant \mathcal{L}\big((x-x')^2 + (y-y')^2\big)^{\alpha/2}$ pour tout $x, y, x', y' \in [0,1]$.

la fonction W^* est constante sur chaque rectangle $[a_{k-1}, a_k[^2$. Alors, l'estimateur $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}$ avec $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0}[K]$ défini par (5.2) satisfait

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant \big(27\sigma^2 + 4b\rho\big)^{1/2}\Big(\frac{3K^2}{n^2} + \frac{2\log K}{n}\Big)^{1/2} + \rho\Big(\frac{2K}{n}\Big)^{1/4}, \tag{5.3}$$

à condition que $\psi_n(\Delta^{(K)}) = \frac{12\log(2e/\Delta^{(K)})}{n\Delta^{(K)}} \leqslant (\sigma/b)^2$.

2. *Supposons que le graphon W^* est α-Hölder continu, c'est-à-dire que $W^* \in \mathbb{H}_{\alpha,\mathcal{L}}$ pour un certain $\alpha \in (0,1]$ et $\mathcal{L} > 0$. Supposons que*

$$\frac{n^{2\alpha+1}}{n\log^4(2n)} \geqslant \mathcal{L}^2\frac{(4b/\sigma)^{4(\alpha+1)} \vee 3}{(25\sigma^2 + 4b\rho)}. \tag{5.4}$$

Soit $\beta = \alpha/(2\alpha+2)$. Alors, il existe un choix de K, n_0 tels que l'estimateur des moindres carrés $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\mathbf{\Theta}}^{\mathsf{LS}}}$ avec $\widehat{\mathbf{\Theta}}^{\mathsf{LS}} = \widehat{\mathbf{\Theta}}^{\mathsf{LS}}_{n_0}[K]$ satisfait

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}}, W^*)^2]^{1/2} \leqslant 6\mathcal{L}^{1-2\beta}\Big(\frac{25\sigma^2 + 4b\rho}{3n^2}\Big)^{\beta} + \Big(\frac{(50\sigma^2 + 8b\rho)\log n}{n}\Big)^{1/2} + \frac{4\mathcal{L}}{n^{\alpha/2}}. \tag{5.5}$$

Nous pouvons également présenter des exemples spécifiques pour les différentes distributions discutées dans la table 5.1, en particulier dans le contexte des graphons Lipschitz. Il convient de noter que pour les graphons constant par morceaux, les résultats obtenus seront les mêmes que ceux présentés dans la table 1.1, avec l'ajout d'un terme d'erreur d'approximation supplémentaire $\rho\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}}\Big)^{1/2}$ comme indiqué dans (1.3). Encore une fois, nous fournissons également une méthode adaptative pour les valeurs inconnues de K et L dans le cas des graphons constants par morceaux.

loi de H_{ij}	Valeurs (σ^2, b)	Condition (5.4)	Borne de risque (5.5)
Bernoulli(ρ)	$(\rho, 1/3)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 n\log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(nm)^{1/4}} + \dfrac{8\sqrt{\rho\log m}}{\sqrt{m}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
Binomial$(N,\rho)/N$	$(\rho/N, 1/3N)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 Nn\log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Nnm)^{1/4}} + \dfrac{8\sqrt{\rho\log m}}{\sqrt{Nm}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
Poisson$(T\rho)/T$	$(\rho/T, 1/3T)$	$\rho^5 \geqslant \dfrac{\mathcal{L}^2 Tn\log^4(2n)}{m^3}$	$\dfrac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Tnm)^{1/4}} + \dfrac{8\sqrt{\rho\log m}}{\sqrt{Tm}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$
sous-Gauss(σ^2)	$(\sigma^2, 0)$	$\sigma^2 \geqslant \dfrac{3\mathcal{L}^2 n\log^4(2n)}{25m^3}$	$\dfrac{11\sqrt{\mathcal{L}\sigma}}{(nm)^{1/4}} + \dfrac{8\sigma\sqrt{\log m}}{\sqrt{m}} + \dfrac{4\mathcal{L}}{\sqrt{m}}$

Table 5.2: Borne supérieure pour les graphons Lipschitz et diverses distributions, dans un cadre non symétrique, avec l'hypothèse supplémentaire que $n \geqslant m$.

Hypothèse de relaxation de l'indépendance A présent, supposons que la matrice $\mathbf{H}$ soit générée en fonction d'un graphon renormalisé W^*, dont les lignes $\boldsymbol{H}_i$ sont indépendantes et

dont la somme vaut 1, comme le prévoit le modèle décrit dans hypothèse 2, où

$$\mathbb{E}[\mathbf{H}|\boldsymbol{U},\boldsymbol{V}] = \boldsymbol{\Theta}^* \quad \text{avec} \quad \Theta^*_{ij} = \frac{W^*(U_i,V_j)}{\sum_{\ell=1}^m W^*(U_i,V_\ell)}\,.$$

Dans le contexte de l'indépendance totale, nous savons déjà que le graphon W^* est non identifiable. De plus, dans ce contexte, nous pouvons multiplier W^* par une constante sans changer la loi de $\mathbf{H}$. Pour remédier à cela, nous définissons une nouvelle classe d'équivalence, où deux graphons W et W' sont considérés équivalents si et seulement s'ils satisfont la relation

$$W = C_x W'(\tau_1 \otimes \tau_2)$$

où C_x est une constante qui pourrait dépendre de la première variable x et τ_1 et τ_2 sont des bijections de $[0,1]$ qui préservent la mesure de Lebesgue. Il est évident que deux tels graphons produiront la même matrice $\mathbf{H}$. Pour la suite, nous supposons que $W^* \in \mathscr{C}$, où

$$\mathscr{C} = \big\{W, I_W(x) = 1/m, \forall x \in [0,1]\big\} \quad \text{with} \quad I_W(x) = \int_0^1 W(x,y),dy.$$

Tout comme précédemment, la distance choisie au sein de la classe $\mathscr{C}$ pour mesurer la qualité des estimateurs est définie comme suit :

$$\begin{aligned}\delta(W',W) &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \left(\iint_{[0,1]^2} \big|W'(\tau_1(u),\tau_2(v)) - W(u,v)\big|^2, du, dv\right)^{\frac{1}{2}} \\ &= \inf_{\tau_1,\tau_2\in\mathcal{M}} \|W' \circ (\tau_1 \otimes \tau_2) - W\|_{\mathbb{L}^2}.\end{aligned}$$

Estimer un graphon devient une tâche difficile en raison du processus complexe de normalisation au sein de la classe $\mathscr{C}$. Dans le théorème suivant, nous présentons une borne supérieure pour l'estimation des graphons morcelés, dans le cadre symétrique.

Théorème 4. *Soit $\mathbf{H} \in [0,1]^{n\times n}$ une matrice aléatoire de taille $n \times n$ qui satisfait l'hypothèse 2 pour un graphon $W^* : [0,1]^2 \to [0,\rho]$. Supposons que chaque ligne de $\mathbf{H}$ somme à un, que sa matric de covariance $\boldsymbol{\Sigma}_i$ satisfait $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} \leqslant \rho_{\boldsymbol{\Sigma}} \leqslant 1$ et que sa moyenne conditionnelle $\boldsymbol{\Theta}^*$ satisfait $\|\boldsymbol{\Theta}^*\|_\infty \leqslant \rho$. Supposons également que le graphon W^* est K-constant par morceaux, ce qui signifie que pour des entiers $K \geqslant 2$ et pour $0 = a_0 < \ldots < a_K = 1$ tels que*

$$\Delta^{(K)} := \min_{k\in[K]} |a_k - a_{k-1}| \geqslant \frac{8\log(nK)}{n}$$

la fonction W^* est constante sur chaque rectangle $[a_{k-1},a_k[^2$. Alors, l'estiamteur $\widehat{W}^{\mathsf{LS}} = W_{\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}}$ avec $\widehat{\boldsymbol{\Theta}}^{\mathsf{LS}} = \widehat{\boldsymbol{\Theta}}^{\mathsf{LS}}_{n_0}[K]$ définit par (5.2) satisfait

$$\mathbb{E}[\delta(\widehat{W}^{\mathsf{LS}},W^*)^2]^{1/2} \leqslant \big(50\rho_{\boldsymbol{\Sigma}} + 6\rho\big)^{1/2}\Big(\frac{3K^2}{n^2} + \frac{2\log K}{n}\Big)^{1/2} + 3\rho\Big(\frac{2K}{n}\Big)^{1/4}$$

pourvu que $\psi_n(\Delta^{(K)}) = \frac{4\log(2e/\Delta^{(K)})}{n\Delta^{(K)}} \leqslant \rho_{\boldsymbol{\Sigma}}$ et $\bar{w} \leqslant \frac{1}{4}e^{0.045n}$ où $\bar{w} = \sum_{\ell=1}^K \frac{1}{w_k}$ and $w_k = a_k - a_{k-1}$.

Le résultat obtenu dans ce théorème est similaire à celui présenté dans le théorème 3,

avec l'exigence supplémentaire que $\bar{w}$ ne soit pas excessivement grand, garantissant ainsi que la taille des intervalles $[a_k, a_{k+1}[$ soit suffisamment grande. Par exemple, si $n = 315$, la condition sur $\bar{w}$ est satisfaite tant que la différence minimale entre les valeurs consécutives de a_k est supérieure ou égale à 10^{-3}. Cette condition apparaît en raison de la normalisation susmentionnée dans la classe $\mathscr{C}$.

5.2.3 Bonre inférieure sur le risque minimax

Dans cette section, nous établissons l'optimalité de l'estimateur des moindres carrés $\widehat{W}^{\mathsf{LS}}$, sous l'hypothèse 1 de pleine indépendance, en démontrant sa vitesse de convergence dans le pire des cas sur la classe $\mathcal{W}_\rho[K, L]$. Cette classe comprend les graphons W qui sont constants sur les intervalles I_k et J_ℓ, où $I_k = [a_k, a_{k+1})$ et $J_\ell = [b_\ell, b_\ell + 1)$ forment une partition de $[0, 1)$.

Nous nous concentrons sur la démonstration de la borne inférieure pour le modèle binomial, mais les techniques utilisées dans la preuve peuvent être étendues aux autres modèles mentionnés dans l'introduction. Cela établit l'optimalité de l'estimateur des moindres carrés dans cette classe.

Théorème 5. *Supposons que conditionnellement à $(\boldsymbol{U}, \boldsymbol{V})$, les entrées $H_{i,j}$ de la matrice observée $\mathbf{H}$ de taille $n \times m$ sont indépendantes et suivent une distribution binomiale de paramètre $(N, W^*(U_i, V_j))$. Il existe des constantes universelles c et $C > 0$ telles que, pour tout K et L supérieurs à C et satisfaisant $KL \geqslant L \log^2 L + K \log^2 K$, ainsi que pour toute valeur de $\rho > 0$, on ait :*

$$\inf_{\widehat{W}} \sup_{W^*} \mathbb{E}_{W^*} \left[\delta^2(\widehat{W}, W^*)\right]^{1/2} \geqslant c \left[\sqrt{\rho} \left(\frac{KL}{Nnm} \wedge \rho + \frac{1}{N\sqrt{nm}} \wedge \rho \right)^{1/2} + \rho \left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}} \right)^{1/2} \right],$$

où l'infimum est pris sur tous les estimateurs possibles $\widehat{W}$ et le supremum est pris sur tous les $W^* \in \mathcal{W}_\rho[K, L]$.

Dans le contexte symétrique où $n = m$ et $K = L$, cette borne inférieure doit être comparée à (5.3) et semble être optimale en termes de taux, à un facteur $\log K$ près. Les figures 22 et 23 montrent les zones violettes où la borne inférieure est de l'ordre de la borne supérieure pour divers paramètres du modèle, ce qui signifie que notre estimateur est optimal dans le sens du minimax.

Pour être plus précis, dans la Figure 6, nous fixons le paramètre de densité ρ et choisissons les paramètres de cluster K et L de manière à ce que $K/n = L/m = \gamma$. La zone violette dans la figure représente les paires (n, m) où la borne inférieure obtenue à partir de Theorem 5 dépasse la moitié de la borne supérieure donnée dans (2.13). D'autre part, la Figure 7 illustre le même critère, mais avec n et m fixes, tout en faisant varier ρ et γ. Il est à noter que l'estimateur des moindres carrés atteint l'optimalité dans de nombreux cas, même dans des cadres très asymétriques où, par exemple, m est nettement plus grand que n.

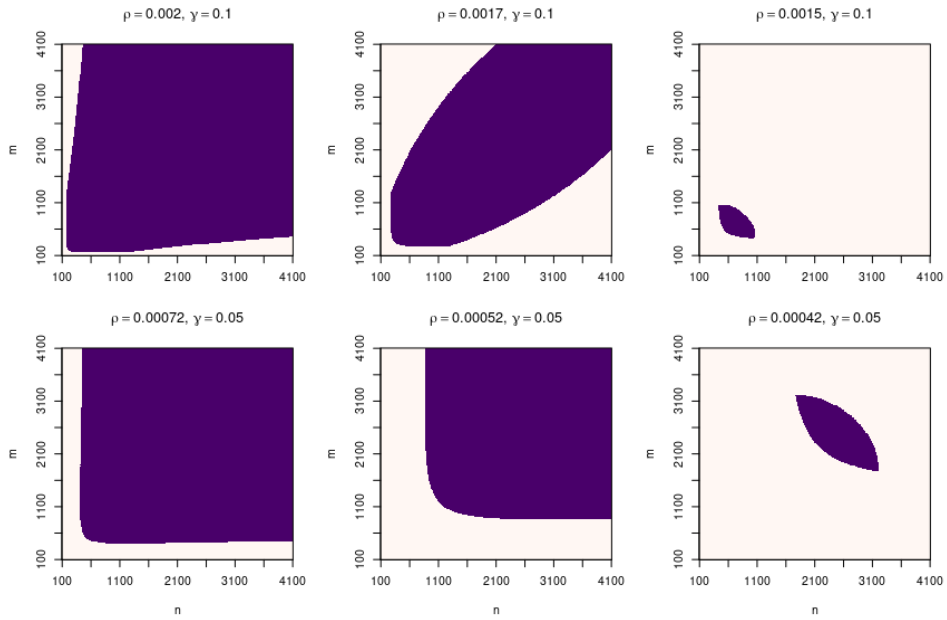


Figure 22: Illustration de l'optimalité de l'estimateur des moindres carrés pour $N = 1$. La zone violette correspond aux valeurs de n et m, pour certaines valeurs fixes de ρ et $\gamma = K/n = L/m$, pour lesquelles la borne inférieure est de l'ordre de la borne supérieure. Plus précisément, lorsque $\rho\gamma^2 \wedge \rho^2 + \rho(nm)^{-1/2} \wedge \rho^2 + 2\rho^2\sqrt{\gamma}$ est supérieur à la moitié de $\rho\gamma^2 + (\rho \log K)/(3m) + (\rho \log L)/(3n) + 2\rho^2\sqrt{\gamma}$. Nous observons que sauf si ρ est très petit, la borne supérieure établie pour l'estimateur des moindres carrés est de l'ordre de la borne inférieure pour tous les estimateurs, pour la plupart des valeurs de n et m.

5.2.4 Algorithme et expériences numériques

L'estimateur des moindres carrés introduit dans l'équation (5.2) et discuté dans les sections précédentes est computationnellement incalculable, c'est à dire qu'il n'est pas possible de calculer cet estimateur en temps polynomial. Dans cette section, notre objectif est de présenter un algorithme qui fournit une approximation computationnellement réalisable de $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$. Bien qu'il n'y ait aucune garantie que l'algorithme produise toujours un estimateur proche de $\widehat{\mathbf{\Theta}}^{\mathsf{LS}}$, on peut s'attendre à ce que ce soit le cas dans de nombreux scénarios.

L'algorithme L'approximation proposée peut être considérée comme une variante de l'algorithme de Lloyd pour le clustering k-means [Llo82]. Pour le décrire, rappelons que l'estimateur des moindres carrés est défini comme une solution qui minimise la distance induite par la norme de Frobenius entre $\mathbf{H}$ et une matrice constante par blocs. Ceci peut être reformulé comme suit :

$$(\widehat{\mathbf{Q}}, \mathbf{Z}_1, \mathbf{Z}_2)^{\mathsf{LS}} \in \arg \min_{\substack{\mathbf{Q} \in \mathbb{R}^{K \times L} \\ \mathbf{Z}_1 \in \mathcal{Z}(n,K,n_0) \\ \mathbf{Z}_2 \in \mathcal{Z}(m,L,m_0)}} \left\| \mathbf{H} - \mathbf{Z}_1 \mathbf{Q} (\mathbf{Z}_2)^\top \right\|_{\mathsf{F}}^2. \tag{5.6}$$

où $\mathbf{Z}_1, \mathbf{Z}_2$ représentent la structure par blocs de la matrice[5] , c'est-à-dire les clusters gauche et droite, et $\mathbf{Q}$ donne les valeurs dans les différents blocs. Il est intéressant de noter que

[5] $\mathcal{Z}(n, K, n_0) = \left\{ \mathbf{Z} \in 0, 1^{n \times K} : \mathbf{Z}\mathbf{1}_K = \mathbf{1}_n \text{ et } \min_{k \in [K]} \mathbf{1}_n^\top \boldsymbol{Z}_{\bullet,k} \geqslant n_0 \right\}$ avec $1_d = (1, \ldots, 1)^\top \in \mathbb{R}^d$.

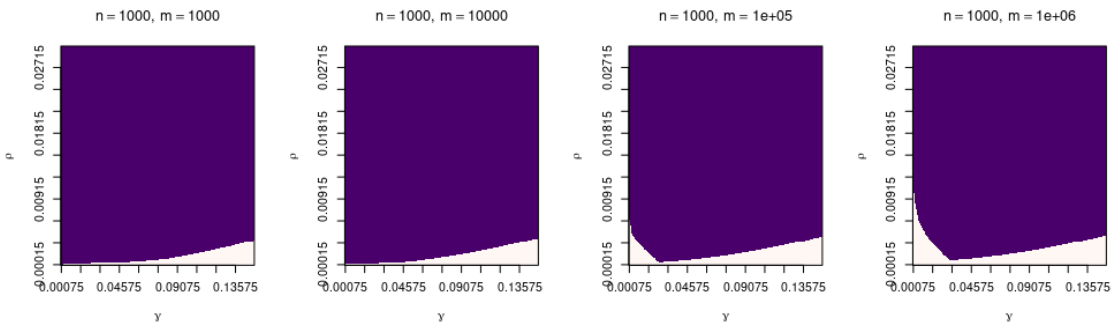


Figure 23: Illustration de l'optimalité de l'estimateur des moindres carrés. La zone violette correspond aux valeurs de ρ et $\gamma = K/n = L/m$, pour certaines valeurs fixes de n et m, pour lesquelles la borne inférieure est de l'ordre de la borne supérieure. Nous observons que sauf si ρ est très petit, la borne supérieure et la borne inférieure sont du même ordre de grandeur.

lorsque deux des trois arguments $\mathbf{Q}$, $\mathbf{Z}_1$ ou $\mathbf{Z}_2$ de la fonction objective sont fixés, le problème de minimisation par rapport à l'argument restant devient possible sur le plan computationnel. Par conséquent, nous pouvons utiliser l'algorithme de minimisation alternée décrit ci-dessous, qui garantit une diminution de la fonction de coût $\mathcal{L}(\mathbf{Z}_1, \mathbf{Q}, \mathbf{Z}_2) = \|\mathbf{H} - \mathbf{Z}_1\mathbf{Q}(\mathbf{Z}_2)^\top\|_{\mathsf{F}}^2$ à chaque itération.

Algorithm 5 Algorithme de minimisation alternée de Lloyd pour l'approximation du LSE (5.6)

Require: $\mathbf{Z}_1$, $\mathbf{Z}_2$ les matrices de clusters à droite et à gauche avec les entrées dans $\{0, 1\}$, $\mathbf{H}$ la matrice des données

Ensure: $(\mathbf{Z}_1, \mathbf{Q}, \mathbf{Z}_2)$ minimum local de $\mathcal{L}(\cdot, \cdot, \cdot)$.

Répéter :

1. Calculer $\mathbf{Q} = (\mathbf{Z}_1^{\text{norm}})^\top \mathbf{H} \mathbf{Z}_2^{\text{norm}}$ où $\mathbf{Z}_1^{\text{norm}}$ est la matrice $\mathbf{Z}_1$ avec des colones renormalisée par rapport à la norm ℓ^1 (le nombre de 1 sur la colonne), et de même pour $\mathbf{Z}_2^{\text{norm}}$.
2. Mettre à jour $\mathbf{Z}_1$ qui minimise $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}_2)$
3. Mettre à jour$\mathbf{Z}_2$ qui minimise $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z_1}, \mathbf{Q}, \mathbf{Z})$

La procédure d'initialisation Comme le montre la figure 24, les matrices initiales choisies pour l'algorithm 5 peuvent influencer de manière significative le résultat final. Une approche pour atténuer ce problème est d'exécuter plusieurs instances de l'algorithme en parallèle, chacune avec différentes matrices d'initialisation choisies de manière aléatoire. Parmi les N estimateurs résultants, l'estimateur final est sélectionné comme celui qui minimise la fonction objective $\mathcal{L}$.

Une autre stratégie, souvent utilisée en conjonction avec l'algorithme de Lloyd, est l'initialisation spectrale. Dans le cas où le graphon est constant par morceaux, le problème peut être vu comme un SBM pour les réseaux bipartites. Une manière d'obtenir des valeurs initiales $(\mathbf{Z}_1, \mathbf{Z}_2)$ est à travers la méthode spectrale proposée dans [ZA19a]. Cette méthode consiste à calculer la décomposition en valeurs singulières tronquées à l'ordre K d'une version régularisée de la matrice $\mathbf{H}$. Les vecteurs singuliers gauches tronqués à l'ordre K sont ensuite utilisés en

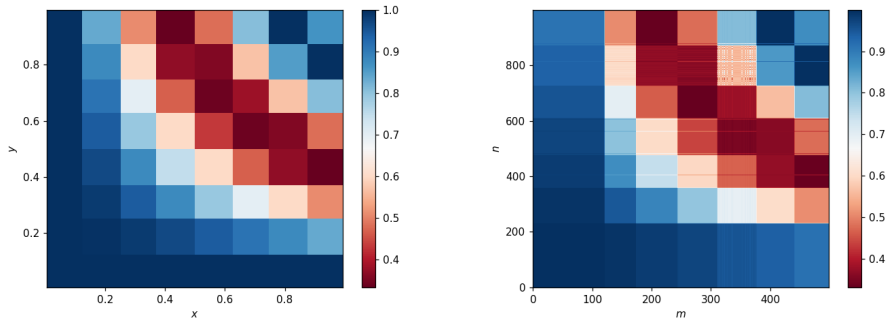
entrée pour le clustering par la méthode des k-mens, ce qui donne une initialisation pour $\mathbf{Z}_1$. Une procédure similaire est appliquée pour obtenir l'initialisation de $\mathbf{Z}_2$. Cette approche d'initialisation spectrale peut fournir un bon point de départ pour l'algorithme de Lloyd et améliorer la qualité de l'estimateur final.

Pour évaluer l'impact de la procédure d'initialisation sur l'estimateur fourni par l'algorithm 5, nous avons effectué plusieurs exécutions en utilisant différentes matrices de clusters. Plus précisément, nous avons utilisé des matrices obtenues par regroupement spectral comme mentionné précédemment, des matrices générées aléatoirement une fois, et des clusters oracles calculés à partir des variables aléatoires non observées. Nous avons ensuite représenté les estimateurs résultants après avoir réorganisé les lignes et les colonnes en fonction de permutations qui ordonnaient des séquences des variables inconnues U_i et V_j. Les résultats, illustrés dans la Figure 24, démontrent clairement que l'initialisation par regroupement donne de meilleurs résultats par rapport à une initialisation aléatoire unique.

Expériences numériques Dans ce paragraphe, nous présentons brièvement quelques expériences numériques pour examiner le comportement de l'erreur d'estimation du graphon $\widehat{W}^{\mathsf{LS}}$ et sa relation avec divers paramètres du modèle. Nous renvoyons le lecteur à la section section 2.6 pour plus de précisions sur ces expériences numériques. Nous commençons par étudier le cas des graphons constant par morceaux et analysons l'erreur d'estimation de la matrice $\mathbf{\Theta}^*$. Nous examinons comment cette erreur varie en fonction du paramètre n pour différentes valeurs de (ρ, K, L), en supposant que $m = n/2$. Les résultats sont représentés dans la figure 25, où les valeurs du graphon W^* sont générées de manière aléatoire, et l'erreur est tracée en échelle logarithmique pour différents types d'initialisation déjà mentionnés dans le paragraphe précédent.

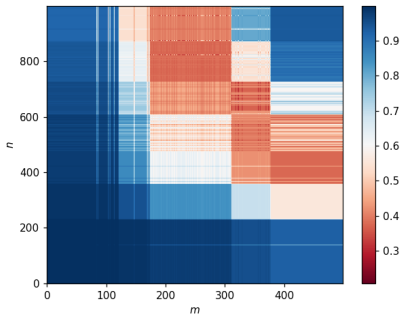
À partir des résultats expérimentaux, on peut observer que l'erreur de la version "spectrale" diminue à mesure que la valeur de n augmente. De plus, elle converge vers l'erreur oracle à un rythme plus rapide lorsque le paramètre de parcimonie ρ est plus élevé et lorsque les ratios n/K et m/L sont plus élevés. Cela est conforme à notre intuition, car un ρ plus élevé implique plus de liens dans le réseau, ce qui conduit à une estimation plus précise. De même, des valeurs plus élevées de n/K et m/L contribuent également à une meilleure précision de l'estimation. En revanche, la version "aléatoire" de l'algorithme présente un comportement plus erratique. Dans la plupart des cas, son erreur dépasse celle de la version "spectrale" lorsque n/K et m/L atteignent un certain seuil.

De plus, dans la figure 26, nous présentons les résultats d'estimation pour un graphon Lipschitz-continu, où les paramètres K et L sont choisis en fonction de n et m respectivement, en suivant les recommandations fournies par nos résultats théoriques. De manière intéressante, et quelque peu surprenante, l'initialisation aléatoire se comporte aussi bien que l'initialisation spectrale. Nous n'avons pas d'explication pour cette observation à ce stade.

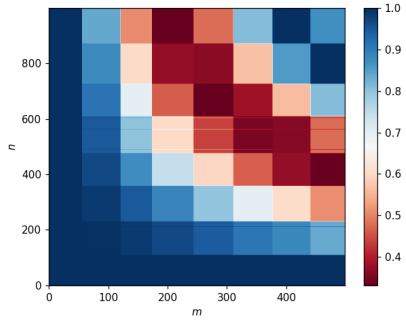


(a) Graphon W^*

(b) Initialisation par clustering



(c) Initialisation aléatoire ($N = 1$)



(d) Initialisation oracle

Figure 24: Illustration de la sensibilité de l'algorithm 5 à la procédure d'initialisation. Nous avons exécuté l'algorithm 5 pour différentes procédures d'initialisation et avons représenté les estimateurs résultants après réarrangement. L'initialisation oracle signifie que nous utilisons les vraies matrices de clusters (inconnues) comme initialisation. Les paramètres choisis ici sont $(\rho, K, L, n, m) = (1, 8, 8, 1000, 500)$.

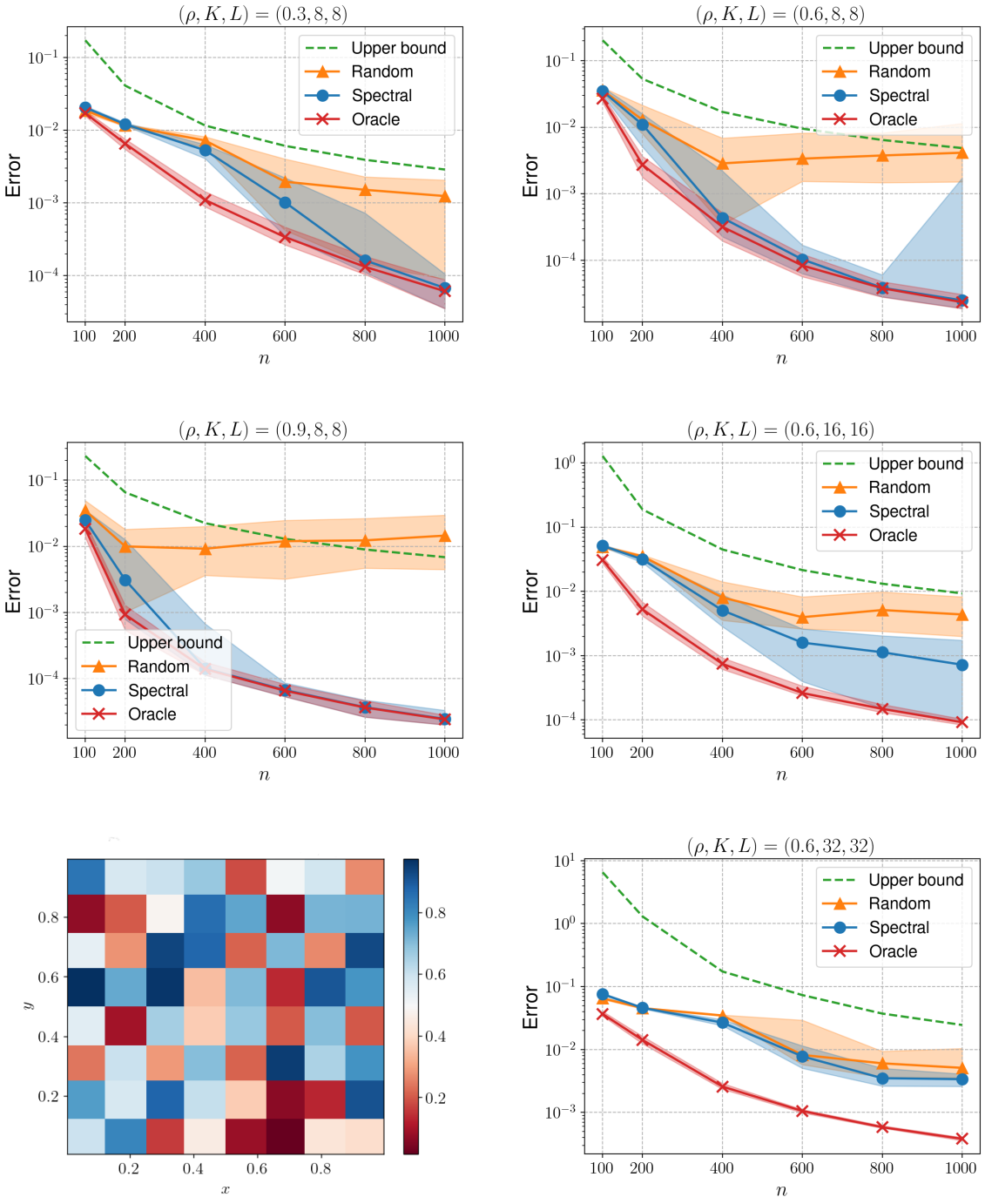
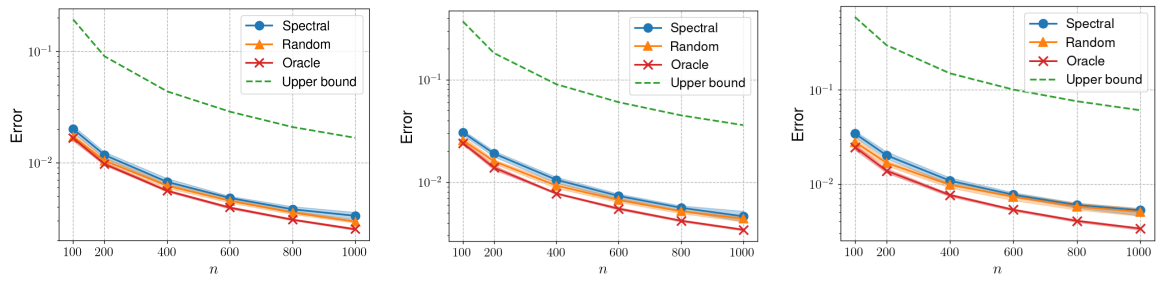
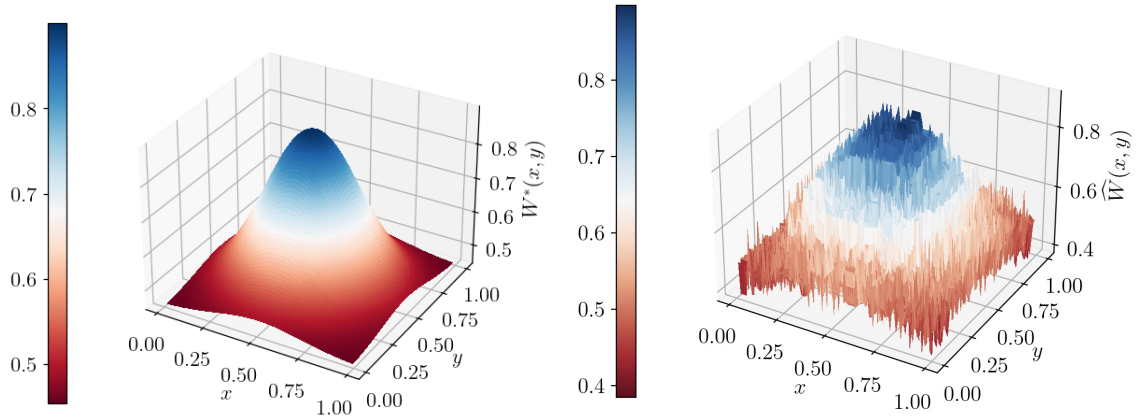


Figure 25: Évolution de l'erreur d'estimation en fonction de n, avec $m = n/2$ pour un graphon aléatoire constant par morceaux pour différentes valeurs de (ρ, K, L).



(a) Error for $\rho = 0.3$ (b) Error for $\rho = 0.6$ (c) Error for $\rho = 0.9$



(d) True graphon W^* (e) Estimated graphon $\widehat{W}$

Figure 26: Évolution de l'erreur d'estimation en fonction de n, avec $m = n/2$ pour un graphon Lipschitz, pour différentes valeurs de ρ. Les courbes représentent l'erreur pour diverses initialisations de l'algorithme de Lloyd (algorithm 5). Le vrai graphon est représenté dans la figure 26d et la figure 26e est une représentation du graphon estimé réarrangé avec une initialisation spectrale. En pratique, nous ne pouvons pas réarranger le graphon estimé car cela nécessite la connaissance des variables latentes.

Bibliography

[Abb18] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[ABH16] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory*, 62(1):471–487, 2016.

[ACC13] Edoardo M. Airoldi, Thiago B. Costa, and Stanley H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neurips 2013*, pages 692–700, 2013.

[Ald81] David J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981.

[BT97] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

[CDP12] Alain A. Celisse, Jean-Jacques J.-J. Daudin, and Laurent L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* , 6:1847–1899, 2012. AMS 2000 subject classifications: Primary 62G05, 62G20; secondary 62E17, 62H30.

[CLC+21] Changxiao Cai, Gen Li, Yuejie Chi, H. Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49(2):944 – 967, 2021.

[CLX18] Yudong Chen, Xiaodong Li, and Jiaming Xu. Convexified modularity maximization for degree-corrected stochastic block models. *Ann. Statist.*, 46(4):1573–1602, 2018.

[CO10] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

[CRV15] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423. PMLR, 2015.

[Dal20] Arnak S. Dalalyan. Exponential weights in multivariate regression and a low-rankness favoring prior. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(2):1465 – 1483, 2020.

[Dal22] Arnak S. Dalalyan. Simple proof of the risk bound for denoising by exponential weights for asymmetric noise distributions. Preprint, Arxiv, December 2022.

[DDG21a] Laurant Davezies, Xavier D'haultfoeuille, and Yannick Guyonvarch. Empirical process results for exchangeable arrays. *The Annals of Statistics*, 49(2):845—-862, December 2021.

[DDG21b] Laurent Davezies, Xavier D'Haultfœuille, and Yannick Guyonvarch. Empirical process results for exchangeable arrays. *The Annals of Statistics*, 49(2):845 – 862, 2021.

[DG14a] Arnaud Dupuy and Alfred Galichon. Personality traits and the marriage market. *Journal of Political Economy*, 122(6):1271–1319, 2014.

[DG14b] Arnaud Dupuy and Alfred Galichon. Personality traits and the marriage market. *Journal of Political Economy*, 122(6):1271–1319, 2014.

[DH73] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

[DMDK+23] Etienne Donier-Meroz, Arnak S. Dalalyan, Francis Kramarz, Philippe Choné, and Xavier D'Haultfoeuille. Graphon estimation in bipartite graphs with observable edge labels and unobservable node labels, 2023.

[dPRST18] Áureo de Paula, Seth Richards-Shubik, and Elie Tamer. Identifying preferences in networks with bounded degree. *Econometrica*, 86(1):263–288, 2018.

[DT07] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007.

[DT08] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[DT12] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012.

[Dze19] Andreas Dzemski. An Empirical Model of Dyadic Link Formation in a Network with Unobserved Heterogeneity. *The Review of Economics and Statistics*, 101(5):763–776, 12 2019.

[FP16a] Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. *29th Annual Conference on Learning Theory*, PMLR 49:943–959, 2016.

[FP16b] Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. In *Proceedings of COLT 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 943–959. JMLR.org, 2016.

[FPV15] Vitaly Feldman, Will Perkins, and Santosh S. Vempala. Subsampled power iteration: a unified algorithm for block models and planted CSP's. In *Advances in Neurips 2015*, pages 2836–2844, 2015.

[GK21] Solenne Gaucher and Olga Klopp. Optimality of variational inference for stochastic block model with missing links. *NeurIPS*, 2021.

[GLMZ16] Chao Gao, Yu Lu, Zongming Ma, and Harrison H. Zhou. Optimal estimation and completion of matrices with biclustering structures. *J. Mach. Learn. Res.*, 17:161:1–161:29, 2016.

[GLZ15] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

[GM21] Chao Gao and Zongming Ma. Minimax Rates in Network Analysis: Graphon Estimation, Community Detection and Hypothesis Testing. *Statistical Science*, 36(1):16 – 33, 2021.

[GMZZ17] Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(60):1–45, 2017.

[GN03] Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473, 2003. Biometrics.

[Gra17] Bryan S. Graham. An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033–1063, 2017.

[Gra20] Bryan S. Graham. Sparse network asymptotics for logistic regression. *Journal of Multivariate Analysis*, October 2020.

[GV19] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

[HLL83] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[JW96] Matthew O. Jackson and Asher Wolinsky. A strategic model of social and economic networks. *J. Econom. Theory*, 71(1):44–74, 1996.

[KTV17] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316 – 354, 2017.

[KV19] Olga Klopp and Nicolas Verzelen. Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, 174:1033–1090, 2019.

[LB06] G. Leung and A.R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

[Lei16] Jing Lei. A goodness-of-fit test for stochastic block models. *Ann. Statist.*, 44(1):401–424, 2016.

[Llo82] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[LM19] Léa Longepierre and Catherine Matias. Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model. *Electron. J. Statist.*, 13(2):4157–4223, 2019.

[LR15] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 2015.

[LZ16] Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

[Neu18] Stefan Neumann. Bipartite stochastic block models with tiny clusters. In *Advances in NeurIPS 2018*, pages 3871–3881, 2018.

[NS01] Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[NST22] Mohamed Ndaoud, Suzanne Sigalla, and Alexandre B. Tsybakov. Improved clustering algorithms for the bipartite stochastic block model. *IEEE Trans. Inf. Theory*, 68(3):1960–1975, 2022.

[OW14a] Sofia C Olhede and Patrick J Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014.

[OW14b] Sophia C. Olhede and Patrick J. Wolfe. Network histograms and universality ofblockmodel approximation. *PNAS*, 11(41):14722–14727, Oct 2014.

[PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[ST07] Daniel A. Spielman and Shang-Hua Teng. Spectral partitioning works: planar graphs and finite element meshes. *Linear Algebra Appl.*, 421(2-3):284–305, 2007.

[Tsy08] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

[vH16] Ramon van Handel. *Probability in High Dimension*. APC 550 Lecture Notes, Princeton University, December 2016.

[vL07] Ulrike von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

[WB17] Y. X. Rachel Wang and Peter J. Bickel. Likelihood-based model selection for stochastic block models. *Ann. Statist.*, 45(2):500–528, 2017.

[WO13] Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation, 2013.

[XJL20] Min Xu, Varun Jog, and Po-Ling Loh. Optimal rates for community estimation in the weighted stochastic block model. *Ann. Statist.*, 48(1):183–204, 2020.

[ZA19a] Zhixin Zhou and Arash A. Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20:1–47, February 2019.

[ZA19b] Zhixin Zhou and Arash A. Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *J. Mach. Learn. Res.*, 20:47:1–47:47, 2019.

[ZA20] Zhixin Zhou and Arash A. Amini. Optimal bipartite network clustering. *J. Mach. Learn. Res.*, 21:40:1–40:68, 2020.

[ZLZ12] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.*, 40(4):2266–2292, 2012.

[ZZ16] Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *Ann. Statist.*, 44(5):2252–2280, 2016.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre: Estimation de graphon pour les graphes bipartites

Mots clés: réseaux bipartite, clustering, algorithme tractable, bornes non asymptotiques

Résumé: De nombreux ensembles de données peuvent être représentés sous forme d'une matrice dont les entrées représentent les interactions entre deux entités de natures différentes. Ces matrices sont appelées matrices d'adjacence de graphes bipartites. Dans notre travail, nous faisons l'hypothèse que ces interactions sont déterminées par des variables latentes non observables.

Dans un premier temps, notre objectif est d'estimer l'espérance conditionnelle de la matrice de données sachant les variables non observables, en supposant que les entrées de la matrice sont i.i.d. Ce problème peut être formulé comme l'estimation d'une fonction bivariée appelée graphon. Dans notre étude, nous nous concentrons sur deux cas, les graphons constants par morceaux et les graphons Hölder. Nous démontrons des bornes de risque pour l'estimateur des moindres carrés, et nous proposons une adaptation de l'algorithme de Lloyd pour calculer une approximation de cet estimateur et nous présentons les résultats d'expériences numériques pour évaluer les performances de ces méthodes.

Dans un deuxième temps, nous abordons les limites du cadre précédent, qui peut ne pas être adapté pour modéliser des situations avec des degrés de sommet bornés. Par conséquent, nous étendons notre étude à l'hypothèse de l'indépendance relaxée, où seules les lignes de la matrice d'adjacence sont supposées indépendantes. Dans ce contexte, nous nous concentrons spécifiquement sur les graphons constants par morceaux.

Title: Graphon estimation in bipartite networks

Keywords: bipartite networks, clustering, tractable algorithm, non asymptotic bounds

Abstract: Many real-world datasets can be represented as matrices where the entries represent interactions between two entities of different natures. These matrices are commonly known as adjacency matrices of bipartite graphs. In our work, we make the assumption that these interactions are determined by unobservable latent variables.

Firstly, our main objective is to estimate the conditional expectation of the data matrix given the unobservable variables under the assumption that matrix entries are i.i.d. This estimation problem can be framed as estimating a bivariate function known as a graphon. In our study, we focus on two cases: piecewise constant graphons and Hölder-continuous graphons. We derive finite sample risk bounds for the least squares estimator. Additionally, we propose an adaptation of Lloyd's algorithm to compute an approximation this estimator and provide results from numerical experiments to evaluate the performance of these methods.

Secondly, we address the limitations of the previous framework, which may not be suitable for modeling situations with bounded degrees of vertices, among other scenarios. Therefore, we extend our study to the relaxed independence assumption, where only the rows of the adjacency matrix are assumed to be independent. In this context, we specifically focus on piecewise constant graphons.

Institut Polytechnique de Paris

91120 Palaiseau, France