

Graphon Estimation in bipartite graphs with observable weights and unobservable node labels

ETIENNE DONIER-MEROZ, ARNAK S. DALALYAN^a, FRANCIS KRAMARZ,
PHILIPPE CHONÉ and and XAVIER D’HAULTFOEILLE

CREST, ENSAE Paris, Institut Polytechnique de Paris, ^aarnak.dalalyan@ensae.fr

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 (the choice of a product by a customer, the 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 paper 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.

MSC2020 subject classifications: Primary 62G05; secondary 62C20

Keywords: graphon estimation; bipartite graph; risk bound; minimax rate

1 Introduction

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

- $U_1, \dots, U_n, V_1, \dots, V_m$ are independent and drawn from the uniform distribution $\mathcal{U}([0, 1])$.
- conditionally to $(\mathbf{U}, \mathbf{V})$, the entries $H_{i,j}$ are independent and $\mathbb{E}[H_{i,j} | \mathbf{U}, \mathbf{V}] = W^*(U_i, V_j)$.

The aforementioned setting corresponds to the practical situation in which there are n customers and m products. Each customer has an unobserved latent feature U and each product has an unobserved latent feature V . We observe the label H that characterizes the interaction between the customer and the product (typically, that the customer chose the product). The function W^* corresponds to the mean value of the interaction for given values of the latent features.

The scope of applicability for both the aforementioned framework and the results derived in this paper encompasses the following main examples:

1. The entries $H_{i,j}$ take the values 0 or 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

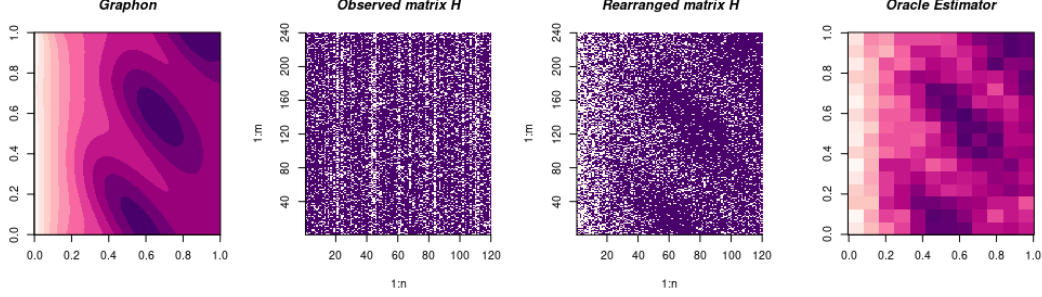


Figure 1: An illustration of the graphon problem. The leftmost graph represents the unknown graphon W^* . The second leftmost graph is the adjacency matrix observed in the graph where the links are made according to the Bernoulli model. The third graph is the adjacency matrix that would be obtained after 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.

- 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. The provided data is composed of an aggregation of N instances of the customer-product graph outlined in the preceding paragraph. Each entry $H_{i,j}$ signifies the observed frequency of customer i purchasing product j . For instance, in a scenario where the customer visited the store $N = 10$ times and $H_{i,j} = 3/10$, it indicates that the customer acquired product i on three out of ten occasions. In this context, the range of values for $H_{i,j}$ spans from 0 to 1. Assuming independence across the aforementioned N trials, the variance of $H_{i,j}$ scales with the order of $1/N$. As a result, this framework exhibits the distinctive trait of having low-noise variance. This particular instance can be encompassed within a broader context of a sub-Gaussian distribution characterized by a small variance parameter σ^2 , which corresponds to $1/N$.
 3. The provided data corresponds to observations collected within a time window of duration $T > 0$. Each $H_{i,j}$ represents the average occurrences of customer i purchasing product j over the time interval with a length of $T > 0$. This scenario can be expressed using the formula $H_{i,j} = N_{i,j}(T)/T$, where $N_{i,j}(T)$ denotes a Poisson random variable characterized by an intensity proportional to T . The parameter T may be large, thereby facilitating accurate estimation. It is important to distinguish this example from the prior one in which the values of $H_{i,j}$ were confined to the interval $[0, 1]$. In the current context, the values of $H_{i,j}$ are not subject to such constraints and can vary arbitrarily in magnitude.

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, and we are also interested in the dependence on the smoothness of W^* , the “sparsity” of interactions denoted by ρ , and the noise level denoted by σ . Specifically, ρ and σ are such that

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

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

In the process of estimating the graphon W^* , an important intermediate step involves estimating the matrix $\Theta^* = W^*(U_i, V_j)$. This matrix estimation is of significant interest in its own right. We approach this task by solving a least-squares problem over the set of block-constant matrices, where the blocks are determined by partitions of the rows and columns of $\mathbf{H}$. Furthermore, we demonstrate that the method of aggregation using exponential weights can be employed to ensure adaptivity to the number of blocks. Under the assumption that the graphon is either piecewise constant or α -Hölder-smooth, we establish non-asymptotic risk bounds for the graphon estimator derived from the estimator of Θ^* . These risk bounds are shown to be rate-optimal in the minimax sense across a broad range of regimes.

Figure 1 illustrates the graphon estimation problem in the case where $\mathbf{H}$ represents the adjacency matrix of a bipartite graph, with entries taking values of either 0 or 1. The figure highlights the significant impact of not knowing the latent variables on the recovery of the graphon. Specifically, the adjacency matrix $\mathbf{H}$ depicted in the second leftmost plot contains limited information about W^* , especially when compared to the rearranged adjacency matrix shown in the third plot.

When $\mathbf{U}$ and $\mathbf{V}$ are unknown, the graphon W^* is unidentifiable. To formalize this, let $\mathcal{M}$ denote the set of all automorphisms $\tau : [0, 1] \rightarrow [0, 1]$ 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 and W' are said to be equivalent if $W = W' \circ (\tau_1 \otimes \tau_2)$ for some $\tau_1, \tau_2 \in \mathcal{M}$. Here, we used the notation $\tau_1 \otimes \tau_2$ for the function from $[0, 1]^2$ to $[0, 1]^2$ defined by $(\tau_1 \otimes \tau_2)(u, v) = (\tau_1(u), \tau_2(v))$. It can be verified that matrices $\mathbf{H}$ and $\bar{\mathbf{H}}$, generated by equivalent graphons W^* and $\bar{W}^*$, share the same distribution. This implies that one can at best estimate the equivalence class containing W^* . This motivates the (pseudo)-distance we use in this work to measure the quality of an estimator $\hat{W}$ of W^* , namely:

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

The main contributions of the present paper 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 two types of estimators: the least-squares estimator that minimizes the squared error over block-constant matrices with a fixed number of clusters, and the exponentially weighted aggregate that combines these least-squares estimators. These results are applicable to adjacency matrices of bipartite graphs, where entries are drawn from the Bernoulli distribution. Moreover, the bounds extend to other distributions, including the binomial distribution, scaled Poisson distributions, 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 whose conditional distribution given latent variables is the Bernoulli one, 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 [Holland et al. \(1983\)](#), is perhaps one of the most studied latent structures for

network data, see also (Nowicki and Snijders, 2001) and (Govaert and Nadif, 2003) 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 (Zhao et al., 2012, Chin et al., 2015, Lei and Rinaldo, 2015, Lei, 2016, Zhang and Zhou, 2016, Wang and Bickel, 2017, Chen et al., 2018, Xu et al., 2020) and literature reviews such as (Gao et al., 2017, Abbe, 2018). 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 (Feldman et al., 2015, Florescu and Perkins, 2016, Neumann, 2018, Zhou and Amini, 2019a, 2020, Cai et al., 2021, Ndaoud et al., 2022).

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 weighted edges. Consistency of graphon estimators has been studied in (Airolidi et al., 2013, Wolfe and Olhede, 2013, Olhede and Wolfe, 2014). In the same problem, minimax-rate-optimality of the least squares estimator has been established in (Gao et al., 2015, 2016, Klopp et al., 2017, Klopp and Verzelen, 2019), see also the survey article (Gao and Ma, 2021). To better present our contributions within the current state-of-the-art, it would be beneficial to provide a brief overview of the contents of these papers. Gao et al. (2015) 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. Gao et al. (2016) extended these results to matrices $\mathbf{H}$ with sub-Gaussian entries, some of which might be missing completely at random. However, as discussed in Section 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, (Klopp et al., 2017) established the minimax-optimal rates of estimation for sparsely connected graphs, where $\rho \ll 1$. While Gao et al. (2015, 2016) measured the estimation error using the normalized Frobenius norm of the difference between the estimated matrix and the true one, Klopp et al. (2017) additionally considered the $\mathbb{L}_2$ -distance between the equivalence classes of graphons. In (Klopp and Verzelen, 2019), minimax optimal rates of graphon estimation in the cut distance have been established for unipartite graphs with binary observations. Computational lower bounds in this problem have been obtained in (Luo and Gao, 2024).

The problems studied in this work are closely related to recent developments in econometrics. In particular, as a consequence of (Aldous, 1981, Theorem 1.4), if the matrix $\mathbf{H}$ can be extended to an infinite array that is row- and column-exchangeable, then there exists a function $g^* : [0, 1]^4 \rightarrow \mathbb{R}$ and independent random variables α , $\{U_i\}$, $\{V_j\}$, and $\{\xi_{ij}\}$ uniformly distributed on $[0, 1]$ such that the random matrices $\mathbf{H}$ and $(g^*(\alpha, U_i, V_j, \xi_{ij}); i \in [n], j \in [m])$ have the same distribution. The problem studied in this paper 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 for the function g^* . If, in addition to $H_{i,j}$, we are also given a feature vector $\mathbf{X}_{i,j}$ for each pair of nodes (i, j) , the extended model defined by $g^*(\mathbf{X}_{i,j}, \alpha, U_i, V_j, \xi_{ij})$ can be considered. This approach is adopted, for instance, in (Graham, 2017), where a specific parametric form $g^*(\mathbf{x}, \alpha, u, v, z) = \mathbb{1}(\mathbf{x}^\top \boldsymbol{\beta} + u + v + \log(z/(1-z)))$ is assumed. In this parametric context, the parameter of interest is the vector $\boldsymbol{\beta}$. Similarly, in (Graham, 2020), the regression function $\int_{[0,1]^3} g^*(\mathbf{x}, \alpha, u, v, z) du dv dz$ is assumed to have the parametric form $\exp\{a + \langle \mathbf{x}, \boldsymbol{\beta} \rangle\} / (1 + \exp\{a + \langle \mathbf{x}, \boldsymbol{\beta} \rangle\})$, and the problem of estimating the parameter vector $(a, \boldsymbol{\beta})$ is studied. Asymptotic results, including the law of large numbers and the central limit theorem for exchangeable arrays, have been established in (Davezies et al., 2021).

Notation For an integer $n \geq 1$, 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

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

and we denote by $\|\mathbf{B}\|_F = \sqrt{\langle \mathbf{B}, \mathbf{B} \rangle}$ 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 $\mathbf{B}_{i,\bullet}$ and $\mathbf{B}_{\bullet,j}$ for the i th row and the j th 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 Estimators of the mean matrix and the graphon

In this section, we define the estimators of the mean matrix $\Theta^* = \mathbb{E}[\mathbf{H}|U, V]$ and of the graphon W^* that are investigated in this paper. We focus here on mathematical definitions only; computational and algorithmic properties of these estimators and their tractable approximations are deferred to Section 4.

2.1 Least squares estimator of Θ^*

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

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

The elements of this set can be seen as assignment matrices: each one of the n customers 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 products. Since the n rows of $\mathbf{H}$ correspond to the customers and the m columns of $\mathbf{H}$ correspond to products, the elements of $\mathcal{Z}(n, K, n_0)$ will be denoted by $\mathbf{Z}^{\text{cust}}$ whereas the elements of $\mathcal{Z}(m, L, m_0)$ will be denoted by $\mathbf{Z}^{\text{prod}}$. Matrices $\mathbf{Z}^{\text{cust}}$ and $\mathbf{Z}^{\text{prod}}$ correspond to a biclustering: the clusters of customers are specified by the matrix $\mathbf{Z}^{\text{cust}}$; in the same way, $\mathbf{Z}^{\text{prod}}$ encodes the clusters of products.

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

$$(\hat{\mathbf{Q}}, \hat{\mathbf{Z}}^{\text{cust}}, \hat{\mathbf{Z}}^{\text{prod}})^{\text{LS}} \in \arg \min_{\substack{\mathbf{Q} \in \mathbb{R}^{K \times L} \\ \mathbf{Z}^{\text{cust}} \in \mathcal{Z}(n, K, n_0) \\ \mathbf{Z}^{\text{prod}} \in \mathcal{Z}(m, L, m_0)}} \|\mathbf{H} - \mathbf{Z}^{\text{cust}} \mathbf{Q} (\mathbf{Z}^{\text{prod}})^\top\|_F^2. \quad (2)$$

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

These estimators computed by (2) lead to the block-constant least squares estimator of Θ^* defined by $\hat{\Theta}^{\text{LS}} = \hat{\mathbf{Z}}^{\text{cust}} \hat{\mathbf{Q}} (\hat{\mathbf{Z}}^{\text{prod}})^\top$. One can write $\hat{\Theta}^{\text{LS}}$ in the following alternative way. Let us consider the class of block-constant matrices

$$\mathcal{T} = \mathcal{T}_{n_0, m_0}^{K, L} = \left\{ \Theta = \mathbf{Z}^{\text{cust}} \mathbf{Q} (\mathbf{Z}^{\text{prod}})^\top : (\mathbf{Q}, \mathbf{Z}^{\text{cust}}, \mathbf{Z}^{\text{prod}}) \in \mathbb{R}^{K \times L} \times \mathcal{Z}_{n, K, n_0} \times \mathcal{Z}_{m, L, m_0} \right\}. \quad (3)$$

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

$$\widehat{\Theta}^{\text{LS}} = \widehat{\Theta}_{n_0, m_0}^{\text{LS}}[K, L] \in \arg \min_{\Theta \in \mathcal{T}_{n_0, m_0}^{K, L}} \|\mathbf{H} - \Theta\|_{\text{F}}^2. \quad (4)$$

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

2.2 Aggregation by exponential weights

The least squares estimator $\widehat{\Theta}^{\text{LS}}$ is constant on KL blocks. The number of these blocks, chosen beforehand, is a hyperparameter of the method. If the true matrix Θ^* is far from being blockwise constant on KL blocks, then the quality of estimation by $\widehat{\Theta}_{n_0, m_0}^{\text{LS}}[K, L]$ 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 1 (with exactly the same U and 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 $\widehat{\Theta}_1^{\text{LS}}, \dots, \widehat{\Theta}_s^{\text{LS}}$ least-squares estimators computed by solving (4) for s different values of (K, L, n_0, m_0) . We define

$$\widehat{\Theta}^{\text{EWA}} = \sum_{\ell=1}^s w_{\ell} \widehat{\Theta}_{\ell}^{\text{LS}}, \quad \text{with} \quad w_{\ell} = \frac{\exp\{-\|\mathbf{H}' - \widehat{\Theta}_{\ell}^{\text{LS}}\|_{\text{F}}^2/\beta\}}{\sum_{r=1}^s \exp\{-\|\mathbf{H}' - \widehat{\Theta}_r^{\text{LS}}\|_{\text{F}}^2/\beta\}}, \quad (5)$$

where $\beta > 0$ is a parameter often referred to as the temperature. Since $\mathbf{H}'$ and $\widehat{\Theta}_{\ell}^{\text{LS}}$ are computed from two independent data matrices, they are independent. This independence plays a crucial role in the proofs of the risk bounds for the EWA. The choice of β depends on the nature of the observations and, more specifically, on the distribution of the noise ξ ; see the next section for further details.

2.3 Adaptations in the case of missing values

The estimators of Θ^* 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 following 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 $p = \mathbf{P}(M_{i,j} = 1)$ 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 $\widetilde{H}_{i,j} = H_{i,j}M_{i,j}/p$, for $i \in [n]$ and $j \in [m]$.

Then, to define the least-squares estimator of Θ^* , it suffices to replace $\mathbf{H}$ with $\widetilde{\mathbf{H}}$ in (2). Similar modifications can be made for defining the exponentially weighted aggregate. Note that this strategy has already been successfully applied in (Gao et al., 2016). If we assume that p is known, the entries of matrix $\widetilde{\mathbf{H}}$ are all observable and the conditional expectation of $\widetilde{\mathbf{H}}$ given (U, V) is still Θ^* . Furthermore, $\text{Var}[\widetilde{H}_{i,j}|U_i, V_j] \leq (\sigma^2 + \rho^2(1-p))/p$, where σ^2 is an upper bound on the conditional variance $\text{Var}[H_{i,j}|U_i, V_j]$ and ρ^2 is an upper bound on the squared coefficients $(\Theta_{i,j}^*)^2$.

In the case of unknown probability p , we can replace $\widetilde{H}_{i,j} = H_{i,j}M_{i,j}/p$ by $\widetilde{H}_{i,j}^* = H_{i,j}M_{i,j}/\widehat{p}$, where $\widehat{p} = \frac{1}{nm} \sum_{i,j} M_{i,j}$ is the proportion of observed entries of $\mathbf{H}$. Since $\widehat{p}$ concentrates around p thanks to the

properties of binomial distributions, the estimators of Θ^* based on $\tilde{\mathbf{H}}^*$ are very close to those based on $\mathbf{H}$. It is therefore plausible that they will satisfy similar oracle inequalities. However, this will necessarily require some technical modifications in the proofs and the statements of the main results, which are beyond the scope of the present work.

2.4 Estimating the graphon

Having estimated the matrix Θ^* , the focus shifts to constructing an estimator of the graphon W^* . To this end, for any $n \times m$ matrix Θ , we define the associated graphon $W_\Theta : [0, 1]^2 \rightarrow [0, 1]$ as a piecewise constant function. More precisely, W_Θ is constant on each rectangle $I_i \times J_j = [(i-1)/n, i/n] \times [(j-1)/m, j/m]$ for $(i, j) \in [n] \times [m]$, and is defined by

$$W_\Theta(u, v) = \Theta_{i,j}, \quad \text{for all } (u, v) \in I_i \times J_j. \quad (6)$$

The rationale behind this definition is the following. Let σ_1 and σ_2 be permutations of $[n]$ and $[m]$, respectively, such that the order statistics satisfy $U_{(i)} = U_{\sigma_1(i)}$ and $V_{(j)} = V_{\sigma_2(j)}$ for any $(i, j) \in [n] \times [m]$. When n and m tend to infinity, $U_{(i)}$ and $V_{(j)}$ are well approximated by $(i-1)/n$ and $(j-1)/m$, respectively. If the true graphon W^* is sufficiently smooth, then for every $(u, v) \in I_i \times J_j$,

$$W^*(u, v) = W^*\left(\frac{i-1}{n}, \frac{j-1}{m}\right) \approx W^*(U_{\sigma_1(i)}, V_{\sigma_2(j)}) = \Theta_{\sigma_1(i), \sigma_2(j)}^* = W_{\Theta^*}\left(\frac{\sigma_1(i)-1}{n}, \frac{\sigma_2(j)-1}{m}\right).$$

Defining $\tau_1 : [0, 1] \rightarrow [0, 1]$ as the measure-preserving bijection given by $\tau_1(u) = (\sigma_1(i) - i + u)/n$ whenever $u \in I_i$, and defining τ_2 analogously, the above display yields $W^*(u, v) \approx W_{\Theta^*}(\tau_1(u), \tau_2(v))$, for all $(u, v) \in [0, 1] \times [0, 1]$, or, equivalently, $W^* \approx W_{\Theta^*} \circ (\tau_1 \otimes \tau_2)$. This provides an intuitive justification for the definition (6).

In addition, for an estimated graphon defined by (6), the estimation error can be easily related to the error, measured by the Frobenius norm, in estimating matrix Θ^* . Indeed, one easily checks that $\|W_{\hat{\Theta}} - W_{\Theta^*}\|_{\mathbb{L}_2} = \|\hat{\Theta} - \Theta^*\|_{\mathbb{F}}/\sqrt{nm}$, which leads to

$$\delta(W_{\hat{\Theta}}, W^*) \leq \|W_{\hat{\Theta}} - W_{\Theta^*}\|_{\mathbb{L}_2} + \delta(W_{\Theta^*}, W^*) \leq \frac{\|\hat{\Theta} - \Theta^*\|_{\mathbb{F}}}{\sqrt{nm}} + \delta(W_{\Theta^*}, W^*). \quad (7)$$

We will apply this inequality to both $\hat{\Theta} = \Theta_{\text{LS}}$ and $\hat{\Theta} = \Theta_{\text{EWA}}$. For simplicity, we often write $\hat{W}^{\text{LS}}$ and $\hat{W}^{\text{EWA}}$ instead of $W_{\Theta_{\text{LS}}}$ and $W_{\Theta_{\text{EWA}}}$, respectively. We will focus on two main classes of graphons: piecewise constant graphons and Hölder continuous graphons, respectively defined by

$$\mathcal{W}[K, L] = \left\{ W : [0, 1]^2 \rightarrow \mathbb{R} : \exists \{I_k\}_{k=0}^{K-1}, \{J_\ell\}_{\ell=0}^{L-1} \text{ s.t. } W = \sum_{k=0}^{K-1} \sum_{\ell=0}^{L-1} W(a_k, b_\ell) \mathbb{1}_{I_k \times J_\ell} \right\}, \quad (8)$$

$$\mathcal{H}_{\alpha, \mathcal{L}} = \left\{ W : [0, 1]^2 \rightarrow \mathbb{R} : |W(\mathbf{x}) - W(\mathbf{x}')| \leq \mathcal{L} \|\mathbf{x} - \mathbf{x}'\|^\alpha, \forall \mathbf{x}, \mathbf{x}' \in [0, 1]^2 \right\}, \quad (9)$$

for some $K, L \in \mathbb{N}^*$ and $\alpha \in (0, 1]$, $\mathcal{L} > 0$. Above, in the definition of $\mathcal{W}[K, L]$, $0 = a_0 < \dots < a_K = 1$, $0 = b_0 < \dots < b_L = 1$, $I_k = [a_k, a_{k+1})$ and $J_\ell = [b_\ell, b_{\ell+1})$ for $k = 0, \dots, K-1$ and $\ell = 0, \dots, L-1$.

The decomposition provided by (7) splits the graphon estimation error into two components: the error of estimating the conditional mean matrix Θ^* and the bias of approximating W^* by the piecewise constant function W_{Θ^*} . 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 1. Let $\Theta_{ij}^* = W^*(U_i, V_j)$ for $i \in [n]$ and $j \in [m]$, where $W^* : [0, 1]^2 \rightarrow [A, B]$ for some A, B such that $-\infty \leq A < B \leq +\infty$.

1. (Piecewise constant graphon) Assume that $W^* \in \mathcal{W}[K, L]$ for some numbers $K, L \in \mathbb{N}^*$. Then

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

2. (Hölder continuous graphon) Assume that $W^* \in \mathcal{H}_{\alpha, \mathcal{L}}$ for some $\alpha \in (0, 1]$ and $\mathcal{L} > 0$. Then

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

The proof of this result, provided in (Donier-Meroz et al., 2024, Section 9.1), extends the techniques developed in the proof of (Klopp et al., 2017, Proposition 3.2) to the case of functions that are not necessarily symmetric. As a minor remark, the constants in our results are smaller than those available in the literature.

3 Finite sample risk bounds

We have introduced in Section 2 the least-squares estimator and the exponentially weighted aggregate for estimating the matrix Θ^* , 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). While the proof of Theorem 1 is deferred to Section 8, the proofs of the other theorems and propositions stated in this paper are provided in (Donier-Meroz et al., 2024).

Throughout this work, we assume that the entries of $\mathbf{H}$ satisfy the following modified Bernstein condition.

Definition 1. For $a, b > 0$, we say that a zero-mean random variable ζ satisfies the (a, b) -Bernstein condition, if

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

The class of random variables satisfying the (a, b) -Bernstein condition extends the classes of sub-exponential and sub-Gaussian random variables. Some basic facts about this condition are provided in (Donier-Meroz et al., 2024, Section 10).

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

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

Theorem 1. Let n, m, n_0, m_0, L , and K be positive integers such that $L \geq 2$, $K \geq 2$, $3 \leq n_0 \leq n$, and $3 \leq m_0 \leq m$. Let ρ, b , and σ be nonnegative numbers. Let $\mathbf{H}$ be an $n \times m$ random matrix with independent entries satisfying $\mathbb{E}[H_{i,j}] \in [0, \rho]$ for all $i \in [n]$ and $j \in [m]$, and such that the centered random variables

$(H_{ij} - \mathbb{E}[H_{ij}])$ satisfy the (σ^2, b) -Bernstein condition. Then the least squares estimator $\widehat{\Theta}^{\text{LS}}$ of the mean matrix $\Theta^* = \mathbb{E}[\mathbf{H}]$, defined in (4), satisfies the exact oracle inequality

$$\frac{\mathbb{E}[\|\widehat{\Theta}^{\text{LS}} - \Theta^*\|_{\text{F}}^2]^{1/2}}{\sqrt{nm}} \leq \inf_{\Theta \in \mathcal{T}_{n_0, m_0}^{K, L}} \frac{\|\Theta - \Theta^*\|_{\text{F}}}{\sqrt{nm}} + (17\sigma^2 + 4b\rho + 24b^2\psi_{n, m}(n_0, m_0))^{1/2} r_{n, m}(K, L)$$

where

$$r_{n, m}(K, L) = \left(\frac{3KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n} \right)^{1/2} \quad \text{and} \quad \psi_{n, m}(n_0, m_0) = \frac{\log(en/n_0)}{m_0} + \frac{\log(em/m_0)}{n_0}. \quad (10)$$

Lower bounds on the minimax risk of all possible estimators, showing that the risk bound in Theorem 1 is rate optimal under various regimes, will be stated in Section 5. Let us mention here the fact that in the particular case $b = 0$ corresponding to a sub-Gaussian distribution, the condition $\mathbb{E}[H_{i, j}] \in [0, \rho]$ can be removed and the claim of the theorem remains true.

Theorem 1 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 1 hold.

1. If H_{ij} 's are drawn from any distribution with support $[0, 1]$ with mean $\Theta_{i, j}^* \leq \rho$, then they satisfy the Bernstein condition both with parameters $(\rho, 1/3)$ and $(1/4, 0)$. Therefore,

$$\frac{\mathbb{E}[\|\widehat{\Theta}^{\text{LS}} - \Theta^*\|_{\text{F}}^2]^{1/2}}{\sqrt{nm}} \leq \inf_{\Theta \in \mathcal{T}_{n_0, m_0}^{K, L}} \frac{\|\Theta - \Theta^*\|_{\text{F}}}{\sqrt{nm}} + 4\sqrt{\Gamma} \left(\frac{KL}{nm} + \frac{\log K}{m} + \frac{\log L}{n} \right)^{1/2}$$

where $\Gamma = 1 \wedge (4\rho + \frac{1}{6m_0} \log(en/n_0) + \frac{1}{6n_0} \log(em/m_0))$.

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

$$\frac{\mathbb{E}[\|\widehat{\Theta}^{\text{LS}} - \Theta^*\|_{\text{F}}^2]^{1/2}}{\sqrt{nm}} \leq \inf_{\Theta \in \mathcal{T}_{n_0, m_0}^{K, L}} \frac{\|\Theta - \Theta^*\|_{\text{F}}}{\sqrt{nm}} + 4\sqrt{\Gamma} \left(\frac{KL}{Nnm} + \frac{\log K}{Nm} + \frac{\log L}{Nn} \right)^{1/2}$$

where $\Gamma = 4\rho + \frac{1}{6Nm_0} \log(en/n_0) + \frac{1}{6Nn_0} \log(em/m_0)$.

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

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

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

$$\frac{\mathbb{E}[\|\widehat{\Theta}^{\text{LS}} - \Theta^*\|_{\text{F}}^2]^{1/2}}{\sqrt{nm}} \leq \inf_{\Theta \in \mathcal{T}_{n_0, m_0}^{K, L}} \frac{\|\Theta - \Theta^*\|_{\text{F}}}{\sqrt{nm}} + 4\sqrt{\Gamma} \left(\frac{KL}{Tnm} + \frac{\log K}{Tm} + \frac{\log L}{Tn} \right)^{1/2}$$

where $\Gamma = 4\rho + \frac{1}{6Tm_0} \log(en/n_0) + \frac{1}{6Tn_0} \log(em/m_0)$.

The remainder term appearing in these risk bounds can be interpreted as

$$\text{noise magnitude} \times \left(\frac{\text{"dimension" of the parameter space}}{\text{sample size}} \right)^{1/2},$$

which is a familiar form for statistical risk bounds. Indeed, the quantity $KL + n \log K + m \log L$ corresponds, up to constants, to the logarithm of the covering number of $\mathcal{T}_{n_0, m_0}^{K, L}$, a standard measure of the complexity of the parameter space. Beyond its heuristic value, this interpretation also explains why the above upper bound is optimal, up to a multiplicative constant, under mild regularity conditions.

Note also that the least-squares estimator, for which the risk bounds above are established, does not require knowledge of ρ , σ , or b . This explains, in particular, why the remainder term in the obtained risk bounds does not vanish when $\sigma + \rho = 0$. This may be contrasted with the results in (Gao et al., 2016), where a remainder term linear in $\sigma + \rho$ is obtained for a least-squares estimator that incorporates knowledge of ρ through the constraints of the minimization problem. Our results have the advantage of showing how ρ enters the risk bound even when the estimator is constructed without using it. However, a limitation of our bounds is that, for small values of ρ , they do not match the minimax lower bounds, since smaller risk bounds can be obtained either for the estimator that outputs the zero matrix, or for the constrained least-squares estimator with the constraint $\|\Theta\|_\infty \leq \rho$ (Gao et al., 2016, Klopp et al., 2017).

Remark 1. As mentioned in the introduction, an upper bound similar to the one of Theorem 1 has been established in (Gao et al., 2016), 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 -sub-Gaussianity 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 with $\rho = 1$, as in the second claim of the last corollary. In this scenario, the risk bound of (Gao et al., 2016) includes a remainder term of order $r_{n,m}(K, L)$ since the sub-Gaussian norm of the arithmetic means 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 $N^{-1/2}$. In Section 5, we demonstrate that this upper bound is tight, at least when n and m are of the same order of magnitude.

3.2 Risk bounds for the EWA $\widehat{\Theta}^{\text{EWA}}$

In signal denoising, pioneering work by Leung and Barron (2006) established sharp bounds for the EWA, initially limited to Gaussian noise. Subsequent progress saw extensions to encompass a wider array of noise distributions, as evidenced by works such as (Dalalyan and Tsybakov, 2007, 2008, 2012, Dalalyan, 2020, 2023). The findings presented in this section are derived from those expounded in (Dalalyan, 2023), which, to the best of our knowledge, remain the sole results in the literature applicable to models featuring asymmetric noise distributions, as arise naturally in the Bernoulli and binomial models.

Theorem 2. Let $\mathbf{H}$ an $n \times m$ matrix with independent entries and let $\Theta^* = \mathbb{E}[\mathbf{H}]$. Let $\mathbf{H}'$ be an independent copy of $\mathbf{H}$. Let $\mathfrak{P}$ be a set of quadruplets $(K, L, n_0, m_0) =: \mathfrak{p}$ and let $\widehat{\Theta}^{\text{LS}}[\mathfrak{p}]$ be the least squares estimator defined by (4). Let $\widehat{\Theta}^{\text{EWA}}[\mathfrak{P}]$ be the exponentially weighted aggregate (5) applied to estimators $\{\widehat{\Theta}^{\text{LS}}[\mathfrak{p}] : \mathfrak{p} \in \mathfrak{P}\}$ with some temperature parameter $\beta > 0$. Let $r_{n,m}(K, L)$ be as in (10).

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

$$\frac{\mathbb{E}[\|\widehat{\Theta}^{\text{EWA}} - \Theta^*\|_F^2]^{1/2}}{\sqrt{nm}} \leq \min_{\mathfrak{p} \in \mathfrak{P}} \left\{ \inf_{\Theta \in \mathcal{T}[\mathfrak{p}]} \frac{\|\Theta - \Theta^*\|_F}{\sqrt{nm}} + 9\sqrt{\frac{\rho}{N}} r_{n,m}(K, L) \right\} + \left\{ \frac{8 \log |\mathfrak{P}|}{3Nnm} \right\}^{1/2}$$

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

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

$$\frac{\mathbb{E}[\|\widehat{\Theta}^{\text{EWA}} - \Theta^*\|_F^2]^{1/2}}{\sqrt{nm}} \leq \min_{\mathfrak{p} \in \mathfrak{P}} \left\{ \inf_{\Theta \in \mathcal{T}[\mathfrak{p}]} \frac{\|\Theta - \Theta^*\|_F}{\sqrt{nm}} + 5\sigma r_{n,m}(K, L) \right\} + \sigma \left\{ \frac{\log |\mathfrak{P}|}{nm} \right\}^{1/2}.$$

To prove the first claim of this theorem, it suffices to combine (Dalalyan, 2023, Cor. 4) with Theorem 1. Similarly, the second claim follows from (Dalalyan, 2023, Cor. 2) and Theorem 1. Similar results can be obtained for arbitrary distribution with bounded support and for the Laplace distribution, using Corollary 3 and Corollary 5 from (Dalalyan, 2023), respectively. Unfortunately, we are currently unaware of any result that facilitate the extension of these bounds to encompass the Poisson distribution and the broader class of sub-Gaussian distributions.

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

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

A suitable combination of the risk bounds established in Theorem 1 on the error of estimators of the mean matrix $\Theta^* = \mathbf{E}[\mathbf{H}]$ and of inequality (7), allows us to get risk bounds for the graphon estimators $\widehat{W}^{\text{LS}}$ and $\widehat{W}^{\text{EWA}}$. We will focus on two classes, piecewise constant and Hölder continuous graphons, for which the evaluation of the approximation error is provided in Proposition 1. Obviously, using this strategy makes the term $\inf_{\Theta \in \mathcal{T}} \|\Theta - \Theta^*\|_F$ —the oracle error—appear in the error bound, where $\mathcal{T}$ is the set of constant by-block-matrices defined in (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 Section 3.1 and Section 3.2, we now return to the original framework of random Θ^* and all the expectations comprise integration with respect to the latent variables U and V .

Proposition 2. Let W^* be α -Hölder continuous, i.e., $W^* \in \mathcal{H}_{\alpha, \mathcal{L}}$ for some $\alpha \in (0, 1]$, $\mathcal{L} > 0$, see (9). Let $\Theta_{i,j}^* = W^*(U_i, V_j)$ for $i \in [n]$ and $j \in [m]$. Let $n_0 \geq 2$, $m_0 \geq 2$, $K \leq n/n_0$ and $L \leq m/m_0$ be four integers. Then, the $n \times m$ matrix Θ^* with entries $\Theta_{i,j}$ satisfies

$$\mathbb{E} \left[\inf_{\Theta \in \mathcal{T}} \frac{\|\Theta - \Theta^*\|_F}{\sqrt{nm}} \right]^{1/2} \leq \frac{3\mathcal{L}}{2} \left(\frac{1}{K^\alpha} + \frac{1}{L^\alpha} \right),$$

where $\mathcal{T} = \mathcal{T}_{n_0, m_0}^{K, L}$ is the set of block-constant matrices defined in (3).

We have now all the necessary ingredients to state the main results of this paper, 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) is computed) is fixed. We then state the result for the exponentially weighted aggregate.

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

1. *If the graphon $W^* \in \mathcal{W}[K, L]$, see (8), such that*

$$\Delta^{(K)} := \min_{k \in [K]} |I_{k-1}| \geq \frac{8 \log(nK)}{n}, \quad \Delta^{(L)} := \min_{\ell \in [L]} |J_{\ell-1}| \geq \frac{8 \log(mL)}{m}, \quad (11)$$

$$\frac{6 \log(2e/\Delta^{(K)})}{m\Delta^{(L)}} + \frac{6 \log(2e/\Delta^{(L)})}{n\Delta^{(K)}} \leq (\sigma/b)^2$$

then, the estimator $\widehat{W}^{\text{LS}} = W_{\widehat{\Theta}^{\text{LS}}}$ with $\widehat{\Theta}^{\text{LS}} = \widehat{\Theta}_{n_0, m_0}^{\text{LS}}[K, L]$ defined by (4) satisfies

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

2. *Let $\alpha \in (0, 1]$, $\mathcal{L} > 0$ and $\beta = \alpha/(2\alpha + 2)$. Without loss of generality, assume that $n \geq m$. If the graphon W^* is Hölder continuous $W^* \in \mathcal{H}_{\alpha, \mathcal{L}}$, see (9), and*

$$\frac{m^{2\alpha+1}}{n \log^4(2n)} \geq \mathcal{L}^2 \frac{(4b/\sigma)^{4(\alpha+1)} \vee 3}{(25\sigma^2 + 4b\rho)}. \quad (12)$$

then, there is a choice of K, L, n_0, m_0 such that the least squares estimator $\widehat{W}^{\text{LS}} = W_{\widehat{\Theta}^{\text{LS}}}$ with $\widehat{\Theta}^{\text{LS}} = \widehat{\Theta}_{n_0, m_0}^{\text{LS}}[K, L]$ satisfies

$$\mathbb{E}[\delta(\widehat{W}^{\text{LS}}, W^*)^2]^{1/2} \leq 6\mathcal{L}^{1-2\beta} \left(\frac{25\sigma^2 + 4b\rho}{3nm} \right)^\beta + \left(\frac{(50\sigma^2 + 8b\rho) \log m}{m} \right)^{1/2} + \frac{4\mathcal{L}}{m^{\alpha/2}}. \quad (13)$$

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 variables¹ $H_{i,j}$ (the Bernstein condition is then fulfilled with $\sigma^2 = \rho$, $b = 1/3$), we obtain

$$\mathbb{E}[\delta(\widehat{W}^{\text{LS}}, W^*)^2]^{1/2} \leq 10 \left(\frac{\rho KL}{nm} + \frac{\rho \log K}{3m} + \frac{\rho \log L}{3n} \right)^{1/2} + \rho \left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}} \right)^{1/2}, \quad (14)$$

provided that $\frac{2}{3m\Delta^{(L)}} \log(2e/\Delta^{(K)}) + \frac{2}{3n\Delta^{(K)}} \log(2e/\Delta^{(L)}) \leq \rho$. In the balanced setting $n = m$, $K = L$ and $\Delta^{(K)} = \Delta^{(L)} = 1/K$, the upper bound in (14) simplifies to

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

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

Distr. of H_{ij}	Values (σ^2, b)	Condition (12)	Risk Bound (13)
Bernoulli(ρ)	$(\rho, 1/3)$	$\rho^5 \geq \frac{\mathcal{L}^2 n \log^4(2n)}{m^3}$	$\frac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(nm)^{1/4}} + \frac{8\sqrt{\rho \log m}}{\sqrt{m}} + \frac{4\mathcal{L}}{\sqrt{m}}$
Binomial(N, ρ)/ N	$(\rho/N, 1/3N)$	$\rho^5 \geq \frac{\mathcal{L}^2 N n \log^4(2n)}{m^3}$	$\frac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Nnm)^{1/4}} + \frac{8\sqrt{\rho \log m}}{\sqrt{Nm}} + \frac{4\mathcal{L}}{\sqrt{m}}$
sub-Gauss(σ^2)	$(\sigma^2, 0)$	$\sigma^2 \geq \frac{3\mathcal{L}^2 n \log^4(2n)}{25m^3}$	$\frac{11\sqrt{\mathcal{L}}\sigma}{(nm)^{1/4}} + \frac{8\sigma\sqrt{\log m}}{\sqrt{m}} + \frac{4\mathcal{L}}{\sqrt{m}}$
Poisson($T\rho$)/ T	$(\rho/T, 1/3T)$	$\rho^5 \geq \frac{\mathcal{L}^2 T n \log^4(2n)}{m^3}$	$\frac{11\sqrt{\mathcal{L}}\rho^{1/4}}{(Tnm)^{1/4}} + \frac{8\sqrt{\rho \log m}}{\sqrt{Tm}} + \frac{4\mathcal{L}}{\sqrt{m}}$

Table 1. Upper bound for Lipschitz-continuous graphons and various distributions.

provided that $\rho \geq 2(K/n) \log(2eK)$. The last expression is of the same order as the rate established in (Klopp et al., 2017, 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 (13) 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 U, 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 (13) is of order $n^{\alpha(1-\alpha)/(2\alpha+2)} \leq n^{0.086}$, when n and m are of the same order. From a practical point of view, this deterioration is not significant, since even for $n = 10^9$, $n^{0.086} \leq 6$.

One can also draw the consequences of the second claim of the theorem under various (conditional to U, V) distributions of $H_{i,j}$. For $\alpha = 1$ ($\beta = 1/4$), that is Lipschitz-continuous graphons, conditions (12) and inequality (13) are reported in Table 1.

To close this section, we state the risk bounds that can be obtained by combining Theorem 2 and Theorem 3. To keep the statement simple, only the case of piecewise constant graphon is presented.

Corollary 2. Let $n \leq m$ and let $\mathbf{H}$ be a $n \times m$ random matrix satisfying Assumption 1 with some (K, L) -piecewise constant graphon $W^* : [0, 1]^2 \rightarrow [0, \rho]$. This means that for some integers $K, L \geq 2$ and for $0 = a_0 < \dots < a_K = 1, 0 = b_1 < \dots < b_L = 1$ such that (11) 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 $\widehat{W}^{\text{EWA}}[\mathfrak{P}] = W_{\Theta^{\text{EWA}}[\mathfrak{P}]}$.

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

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

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

2. (Gaussian model) Assume that, conditionally to (U, 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 $|W^*(u, v)| \leq \rho$ and $\sigma^2(u, v) \leq \sigma^2$ for every $u, v \in [0, 1]$. For $\beta = 4\sigma^2$, we have

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

4 Tractable approximation of the least-squares estimator

The least squares estimator introduced in (2) and studied in previous sections is a solution to a combinatorial optimization problem that is computationally intractable. Specifically, 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 to $\widehat{\Theta}^{\text{LS}}$. While there is no guarantee that the proposed algorithm always yields an estimator close to $\widehat{\Theta}^{\text{LS}}$, it is plausible that this holds in many practical cases.

It should also be stressed that, as shown in (Luo and Gao, 2024), there exists a computation-information gap in the problem of graphon estimation. Roughly speaking, this implies that the rate of convergence for the best possible computationally tractable algorithm is provably worse than that of the least squares estimator. For instance, when edge labels are drawn from a Bernoulli distribution and the class of piecewise constant unipartite graphons with K blocks and sample size n is considered, (Luo and Gao, 2024) proved that the best possible rate of convergence for an estimator representable as a low-degree polynomial of the input matrix $\mathbf{H}$ is, up to logarithmic factors, of order $(K/n)^{1/2}$. For large K , this rate is significantly worse than the rate $(K/n) + ((\log K)/n)^{1/2}$ achieved by the least-squares estimator. Although the Lloyd-type algorithm presented below is not a polynomial of $\mathbf{H}$, it is likely that its error rate suffers from a similar degradation.

The approximation to the least-squares estimator proposed in this section can be viewed as an adaptation of Lloyd’s algorithm for k -means clustering (Lloyd, 1982). To describe it, let us recall that the least squares estimator is defined by

$$(\widehat{\mathbf{Q}}, \widehat{\mathbf{Z}}^{\text{cust}}, \widehat{\mathbf{Z}}^{\text{prod}})^{\text{LS}} \in \arg \min_{\substack{\mathbf{Q} \in \mathbb{R}^{K \times L} \\ \mathbf{Z}^{\text{cust}} \in \mathcal{Z}(n, K, n_0) \\ \mathbf{Z}^{\text{prod}} \in \mathcal{Z}(m, L, m_0)}} \|\mathbf{H} - \mathbf{Z}^{\text{cust}} \mathbf{Q} (\mathbf{Z}^{\text{prod}})^{\top}\|_{\text{F}}^2. \quad (15)$$

It turns out that when two of the three arguments $\mathbf{Q}$, $\mathbf{Z}^{\text{cust}}$, and $\mathbf{Z}^{\text{prod}}$ of $\mathcal{L}$ are fixed, the minimization problem with respect to the third becomes tractable. Consequently, we employ an alternating minimization algorithm, which guarantees that the cost function $\mathcal{L}(\mathbf{Z}^{\text{cust}}, \mathbf{Q}, \mathbf{Z}^{\text{prod}}) = \|\mathbf{H} - \mathbf{Z}^{\text{cust}} \mathbf{Q} (\mathbf{Z}^{\text{prod}})^{\top}\|_{\text{F}}^2$ decreases at each iteration. Various versions of this algorithm have been studied in the literature on estimation and detection in the presence of latent structure (Coja-Oghlan, 2010, Lei and Rinaldo, 2015, Lu and Zhou, 2016, Giraud and Verzelen, 2019).

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

Minimization in the second argument $\mathbf{Q}$ When the clusters are known, meaning that we know matrices $\mathbf{Z}^{\text{cust}} \in \mathcal{Z}(n, K, n_0)$ and $\mathbf{Z}^{\text{prod}} \in \mathcal{Z}(m, L, m_0)$, see (1), the solution to

$$\widehat{\mathbf{Q}} = \arg \min_{\mathbf{Q} \in \mathbb{R}^{K \times L}} \mathcal{L}(\mathbf{Z}^{\text{cust}}, \mathbf{Q}, \mathbf{Z}^{\text{prod}})$$

Algorithm 1 Lloyd's algorithm of alternating minimization for approximating the LSE (15)

Require: $\mathbf{Z}^{\text{cust}}, \mathbf{Z}^{\text{prod}}$ the left and right cluster matrices with entries in $\{0, 1\}$, $\mathbf{H}$ the data matrix.

Ensure: $(\mathbf{Z}^{\text{cust}}, \mathbf{Q}, \mathbf{Z}^{\text{prod}})$ local minimizer of $\mathcal{L}(\cdot, \cdot, \cdot)$.

Repeat:

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

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

Minimization with respect to $\mathbf{Z}^{\text{cust}}$ We focus now on the problem of minimizing the cost function $\mathcal{L}(\mathbf{Z}^{\text{cust}}, \mathbf{Q}, \mathbf{Z}^{\text{prod}})$ over $\mathbf{Z}^{\text{cust}} \in \mathcal{Z}(n, K, n_0)$. 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 $\hat{\mathbf{Z}}^{\text{cust}} \in \mathbb{R}^{n \times K}$ that minimizes $\mathbf{Z} \mapsto \mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}^{\text{prod}})$ 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 $C_\ell^{\text{prod}} = \{j \in [m], Z_{j\ell}^{\text{prod}} = 1\}$ to be the ℓ -th right cluster and introduce notation

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

Simple algebra yields $\mathcal{L}(\mathbf{Z}, \mathbf{Q}, \mathbf{Z}^{\text{prod}}) = \sum_{i=1}^n \|H_{i,\bullet} - \mathbf{Z}_{i,\bullet}^\top \mathbf{Q} (\mathbf{Z}^{\text{prod}})^\top\|_2^2$, where $\mathbf{Z}_{i,\bullet}$ is i th row of $\mathbf{Z}$. Since $\mathbf{Z}_{i,\bullet}$ is allowed to have only one nonzero entry, and it should be equal to one, $\mathbf{Z}_{i,\bullet}^\top \mathbf{Q} (\mathbf{Z}^{\text{prod}})^\top$ is merely a row of $\mathbf{Q} (\mathbf{Z}^{\text{prod}})^\top$. This implies that $\hat{\mathbf{Z}}_{ij}^{\text{cust}} = \mathbb{1}(j = \hat{k}_i)$, where

$$\hat{k}_i = \arg \min_{k \in [K]} \|H_{i,\bullet} - \mathbf{Q}_{k,\bullet} (\mathbf{Z}^{\text{prod}})^\top\|_2^2 = \arg \min_{k \in [K]} \sum_{\ell=1}^L \sum_{j \in C_\ell^{\text{prod}}} (H_{ij} - Q_{k\ell})^2.$$

We can rewrite the above expression of $\hat{k}_i$ as $\hat{k}_i = \arg \min_{k \in [K]} \|\mathbf{D}^{1/2} (\bar{\mathbf{H}}_{i,\bullet}^{\text{prod}} - \mathbf{Q}_{k,\bullet})^\top\|_2^2$. Thus, to determine $\mathbf{Z}^{\text{cust}}$, it suffices to compute the matrix $\bar{\mathbf{H}}^{\text{prod}} \in \mathbb{R}^{n \times L}$ and then find for each row of $\bar{\mathbf{H}}^{\text{prod}}$ the closest row of $\mathbf{Q}$. Of course, the same procedure is valid for minimizing $\mathcal{L}$ with respect to $\mathbf{Z}^{\text{prod}}$ for known $\mathbf{Z}^{\text{cust}}$ and $\mathbf{Q}$.

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

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

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

Note that ϕ is linear in $\mathbf{Z}$, whereas $\tilde{\mathcal{Z}}(n, K, n_0)$ is a convex polytope containing $\mathcal{Z}(n, K, n_0)$.

Proposition 3. *The following two claims hold.*

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

The set $\tilde{\mathcal{Z}}(n, K, n_0)$ is a convex polytope because it is defined by linear constraints. A well known result (Bertsimas and Tsitsiklis, 1997, p 65, Thm 2.7) implies that if $\mathcal{L}(\cdot, \mathbf{Q}, \mathbf{Z}^{\text{prod}})$ has a minimizer in the polytope $\tilde{\mathcal{Z}}(n, K, n_0)$, 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, including linear programming (LP) solvers such as the simplex method (Dantzig, 1963) or interior-point methods (Karmarkar, 1984, Mehrotra, 1992), as well as specialized algorithms for vertex enumeration like the double description method (Motzkin et al., 1953) or reverse search (Avis and Fukuda, 1996). These methods are implemented in optimization libraries such as Gurobi (Gur, 2023), CPLEX (IBM, 2023), and CVX (Grant and Boyd, 2014).

Initialization The initialization of Algorithm 1 might have a strong impact on the final result. One possible strategy is to run in parallel N instances of the algorithm with different initial 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 but the bi-stochastic block model for bipartite networks. Therefore, initial values $(\mathbf{Z}^{\text{cust}}, \mathbf{Z}^{\text{prod}})$ can be obtained, for instance, by the spectral method from (Zhou and Amini, 2019b). 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 $\mathbf{Z}^{\text{cust}}$. The procedure is similar for $\mathbf{Z}^{\text{prod}}$.

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 $(\mathbf{Z}^{\text{cust}}, \mathbf{Z}^{\text{prod}})$. 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 6, results are quite satisfactory.

5 Lower bounds on the minimax risk

In this section, we show that the least squares estimator $\hat{W}^{\text{LS}}$ achieves optimality among all possible estimators in terms of its worst-case rate of convergence over the class $\mathcal{W}_\rho[K, L]$ of piecewise constant graphons, as defined below. While obtaining sharp lower bounds for the class of Hölder continuous graphons is more challenging, such an analysis is beyond the scope of the present work.

Let $\{I_k = [a_k, a_{k+1}) : k \in [K]\}$ and $J_\ell = \{[b_\ell, b_{\ell+1}) : \ell \in [L]\}$ form a partition of $[0, 1)$ into intervals.

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

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 4. Assume that conditionally to $(\mathbf{U}, \mathbf{V})$, the entries $H_{i,j}$ of the observed $n \times m$ matrix $\mathbf{H}$ are independent, and each $NH_{i,j}$ is drawn from the binomial distribution with parameter $(N, W^*(U_i, V_j))$.

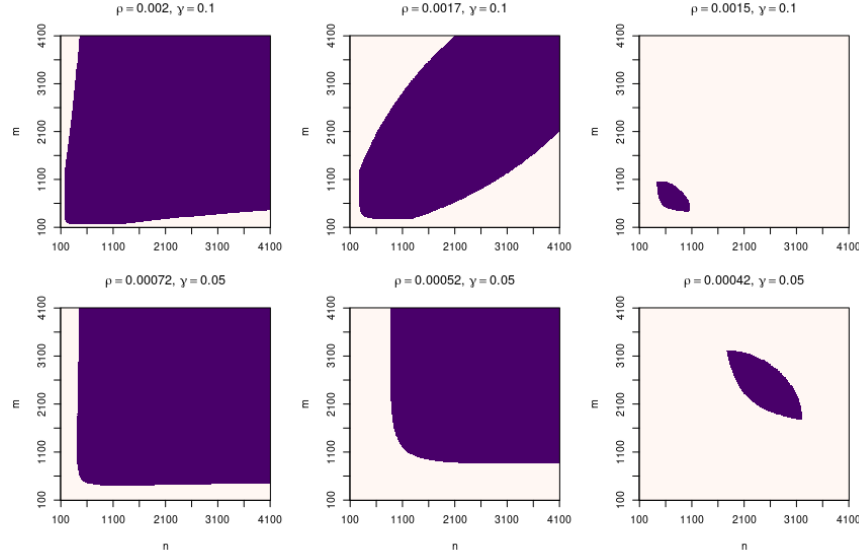


Figure 2: Illustration of the optimality of the least squares estimator for $N = 1$. The purple area corresponds to the values of n and m , for some fixed values of ρ and $\gamma = K/n = L/m$, for which the lower bound is within a constant factor of the upper bound. More precisely, when $\rho\gamma^2 \wedge \rho^2 + \rho(nm)^{-1/2} \wedge \rho^2 + 2\rho^2\sqrt{\gamma}$ 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 .

There exist universal constants $c, C > 0$, such that for any $K, L > C$ satisfying $KL \geq L \log^2 L + K \log^2 K$ and for any $\rho > 0$,

$$\inf_{\hat{W}} \sup_{W^*} \mathbb{E}[\delta^2(\hat{W}, W^*)]^{1/2} \geq c\sqrt{\rho} \left(\frac{KL}{Nnm} \wedge \rho + \frac{1}{N\sqrt{nm}} \wedge \rho \right)^{1/2} + c\rho \left(\sqrt{\frac{K}{n}} + \sqrt{\frac{L}{m}} \right)^{1/2}, \quad (16)$$

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

The right-hand side of (16) should be compared to (14). It can be observed that when the values of n, m, K, L , and ρ are such that the dominating term in the upper bound is either $\rho(K/n)^{1/4}$ or $\rho(L/m)^{1/4}$, the lower bound in (16) matches the upper bound in order. In this case, the LSE of the graphon is minimax-rate-optimal. Similarly, if the dominating term is $(\rho KL/(Nnm))^{1/2}$ and $\rho \geq KL/(Nnm)$, the LSE remains minimax-rate-optimal. It is worth noting that when ρ is very small—specifically, smaller than both $KL/(Nnm)$ and $(N^2nm)^{-1/2}$ —the lower bound in (16) becomes of order ρ , which may be much smaller than the upper bound established for the LSE. This discrepancy is not an artifact of the proof but reflects the fact that in such cases, the naive estimator $\hat{W} \equiv 0$ outperforms the LSE. Moreover, this naive estimator is minimax-rate-optimal under the stated constraint on ρ . Figure 2 illustrates the regions of n and m where the lower and upper bounds are of the same order, thereby confirming the optimality of the LSE. Similarly, Figure 3 depicts the regions in the ρ - γ plane, where $\gamma = K/n = L/m$ (for fixed n and m), in which the lower bound exceeds half of the upper bound. Notably, even for highly unbalanced graphs ($m \gg n$), the purple region almost fully covers the square, indicating that the LSE remains minimax-rate-optimal throughout this region.

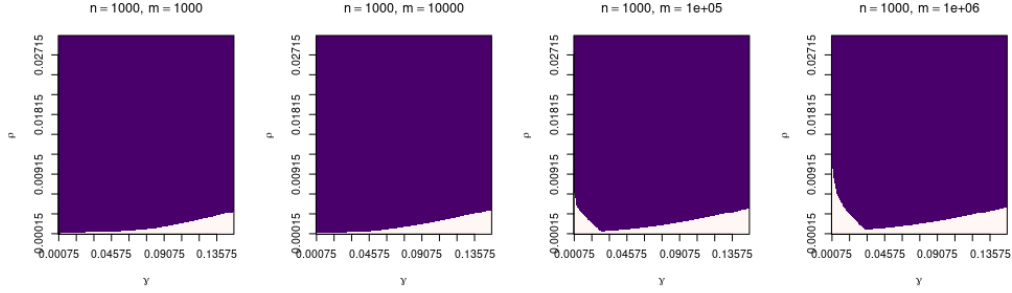


Figure 3: Illustration of the optimality of the least squares estimator. The purple area corresponds to the values of ρ and $\gamma = K/n = L/m$, for some fixed values of n and m , for which the lower bound is within a constant factor of the upper bound. We observe that unless ρ is very small, the upper bound and the lower bound are of the same order.

6 Numerical experiments

In this section, we present the results of some numerical experiments illustrating the behavior of the error of the estimated graphon $\widehat{W}^{\text{LS}}$ and its dependence on different parameters of the model. We first consider the case of piecewise constant graphons and study the estimation error of the matrix Θ^* . 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. In all our experiments, the entries H_{ij} are binary, taking values in $\{0, 1\}$. This corresponds to the conditionally Bernoulli model, or the conditionally binomial model with $N = 1$.

Estimation error of the block-constant matrix Θ^* We report the results of two different experimental set-ups, referred to as rand-graphon and cos-graphon. The two setups 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 uniformly at random between 0 and ρ , while in the cos-graphon set-up, W^* is defined as

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

The results obtained in these two set-ups are depicted in Figure 4 and Figure 5, respectively. In each experiment, we chose $m = n/2$ and computed the median of the squared error $\frac{1}{nm} \|\widehat{\Theta} - \Theta^*\|_F^2$ for 50 independent repetitions. The estimator $\widehat{\Theta}$ was computed by Algorithm 1.

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 1. 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 1, we chose $\gamma = 10^{-3}$ and the maximum number of iterations equal to 40.

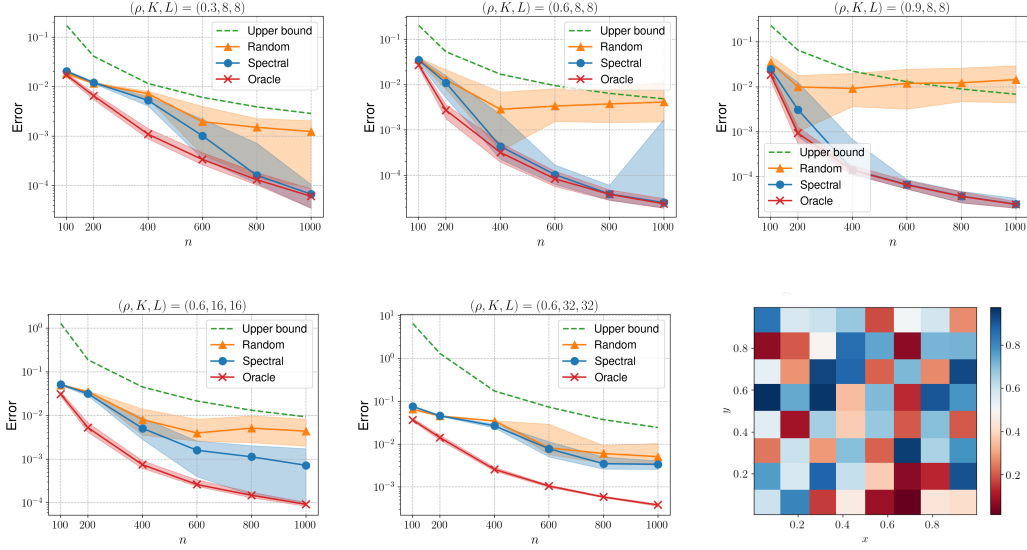


Figure 4: Graphon estimation in the rand-graphon set-up.

The experimental results reveal a clear trend: In the vast majority of cases, the error of the “spectral” version decreases as n increases. Furthermore, when ρ takes on larger values and both n/K and m/L are sufficiently large, the error of the estimator converges to that of the oracle. This behavior aligns with intuition: A larger ρ corresponds to a higher density of links, which facilitates more accurate estimations. Similarly, larger values of n/K and m/L enhance the estimator’s accuracy.

In contrast, the “random” version of the algorithm exhibits more erratic behavior. In most cases, its error exceeds that of the “spectral” version, especially when the ratios n/K and m/L are large. Our empirical study further reveals that, in scenarios with a large number of communities, the performance of the “spectral” version degrades, and its accuracy becomes comparable to that achieved by random initialization.

Estimation error for Hölder-continuous graphons To illustrate the behavior of the estimator of the graphon in the Hölder-continuous case, we consider the function displayed in Figure 6 and given by

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

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 7. 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 $\tilde{\delta}$. Roughly speaking, $\tilde{\delta}$ 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 1. 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 $\tau_{\sigma_1}(u)$ and $\tau_{\sigma_2}(v)$ are given by

$$\tau_{\sigma_1}(u) = \sum_{i=1}^n \left(\frac{\sigma_1(i) - 1}{n} - u - \frac{i - 1}{n} \right) \mathbb{1}_{[i-1, i)}(nu), \quad \tau_{\sigma_1}(1) = \frac{\sigma_1(n)}{n},$$

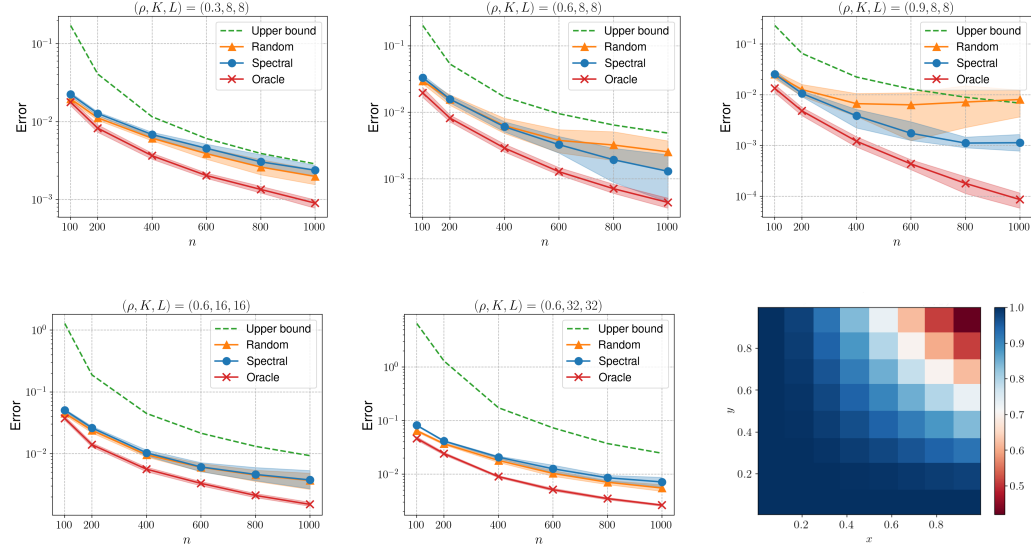


Figure 5: Graphon estimation in the cos-graphon set-up, $W^*(x, y) = \frac{2\rho}{3} + \frac{\rho}{3} \cos(3\pi[Kx][Ly])$.

$$\tau_{\sigma_2}(v) = \sum_{j=1}^m \left(\frac{\sigma_2(j) - 1}{m} - v - \frac{j-1}{m} \right) \mathbb{1}_{[j-1, j)}(nu), \quad \tau_{\sigma_2}(1) = \frac{\sigma_2(m)}{m}.$$

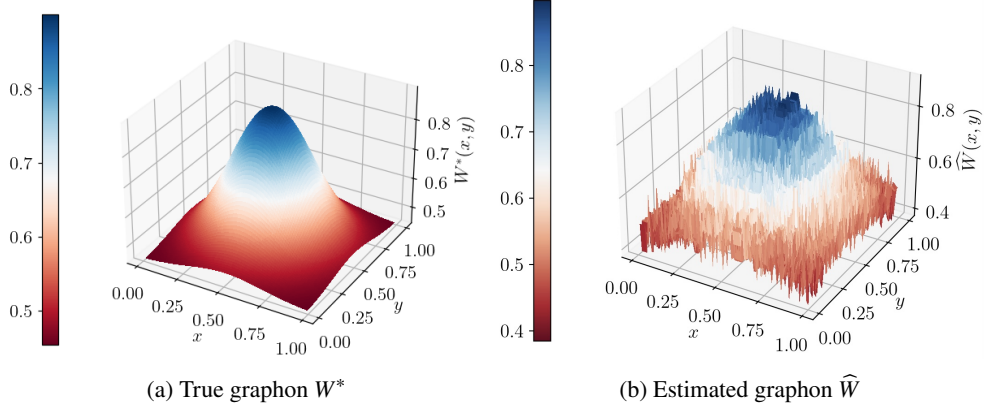
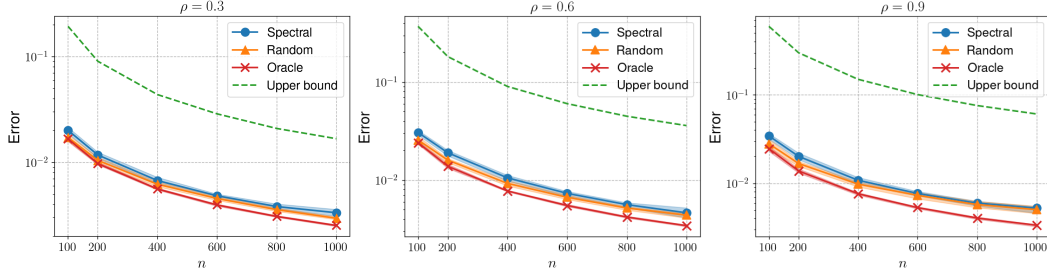
Then, we define $\tilde{\delta}^2(\widehat{W}, W^*) = \|W^* \circ (\tau_{\sigma_1} \otimes \tau_{\sigma_2}) - \widehat{W}\|_{\mathbb{L}_2}^2$. This quantity can also be written as

$$\tilde{\delta}^2(\widehat{W}, W^*) = \|W^*\|_{\mathbb{L}_2}^2 - 2 \sum_{i=1}^n \sum_{j=1}^m \widehat{\Theta}_{ij} \int_{\frac{\sigma_1(i)-1}{n}}^{\frac{\sigma_1(i)}{n}} \int_{\frac{\sigma_2(j)-1}{m}}^{\frac{\sigma_2(j)}{m}} W^*(x, y) dx dy + \|\widehat{\Theta}\|_F^2.$$

In the numerical experiments, the integrals appearing on the right-hand side of the last display are approximated using Riemann sums. Interestingly, the error curves obtained through Monte Carlo simulations closely match the shapes predicted by the theoretical results. Notably, and somewhat unexpectedly, the random initialization performs as well as the spectral initialization. At this stage, we do not have a clear explanation for this observation.

7 Conclusion and Outlook

In this paper, we studied the problem of estimating the conditional mean of a random matrix generated by a bivariate graphon and unobserved latent variables. Our work establishes the fundamental limits of estimation in this setting within a nonparametric framework encompassing both piecewise-constant and Hölder-continuous function classes. Our analysis focused on deriving finite-sample risk bounds for two types of estimators: the least-squares estimator, which minimizes the squared error over block-constant matrices with a fixed number of clusters, and the exponentially weighted aggregate, which adaptively combines these least-squares estimators. These results apply not only to adjacency matrices of bipartite graphs with Bernoulli-distributed entries, but also extend to other distributions, such as binomial, scaled

Figure 6: Graphon representation for $\rho = 0.9$.Figure 7: Evolution of the estimation error for a Lipschitz graphon, $m = n/2$.

Poisson, and sub-Gaussian. We demonstrated that the least-squares estimator achieves rate-optimal performance in the minimax sense across a broad range of regimes, including sparse and dense graphs as well as low-noise settings. The derived risk bounds highlight the dependence on key parameters such as the matrix dimensions n and m , the smoothness of W^* , the sparsity parameter ρ , and the noise level σ .

To address the computational intractability of the least-squares estimator, we proposed an adaptation of Lloyd's algorithm based on alternating minimization and convex relaxation, providing a computationally tractable approximation. It is important to acknowledge, however, the existence of a computation-information gap in graphon estimation, as demonstrated in (Luo and Gao, 2024): the rate of convergence for computationally tractable algorithms, such as those representable as low-degree polynomials of the input matrix $\mathbf{H}$, can be provably worse than that of the least-squares estimator. While our Lloyd-type algorithm is not itself a polynomial of $\mathbf{H}$, it is reasonable to expect that its error rate may suffer from a similar degradation, particularly in regimes where the number of blocks KL is large.

Several important directions remain open for future research. First, the statistical complexity of the Lloyd-type algorithm warrants further investigation. Understanding its convergence properties and the rate of convergence of its risk, in the spirit of (Balakrishnan et al., 2017), for both piecewise-constant and Hölder-continuous graphons are interesting avenues to pursue. Second, an important direction involves relaxing the assumption that two independent data copies are available for constructing an adaptive estimator. While this assumption ensures theoretical tractability in our analysis of the exponentially weighted aggregate, it may not always hold in practice. Exploring ways to relax it, potentially by

integrating suitably chosen complexity penalties into the least-squares framework, is a natural next step. Finally, while our framework assumes that the graphon W^* is defined on $[0, 1]^2$ and that latent variables are uniformly distributed on $[0, 1]$, these assumptions are less restrictive than they may appear. Using the probability integral transform of the latent variables, our setting can be shown to be equivalent to the general framework in which latent variables are drawn from arbitrary continuous distributions. However, this transformation may break the smoothness of the original graphon. Therefore, further exploration of the robustness of our methods to the choice of latent distribution is left to future work.

8 Proof of Theorem 1 (risk bound for $\widehat{\Theta}^{\text{LS}}$)

Let us define $\Pi_{\mathcal{T}}(\Theta^*) = \arg \min_{\Theta \in \mathcal{T}} \|\Theta - \Theta^*\|_F$, the best approximation of Θ^* in Frobenius norm by a block-constant matrix. The matrix $\Pi_{\mathcal{T}}(\Theta^*)$ has at most KL distinct entries each of which is the mean of the entries of a submatrix of Θ^* . Since $\widehat{\Theta} = \widehat{\mathbf{Z}}^{\text{cust}} \widehat{\mathbf{Q}} (\widehat{\mathbf{Z}}^{\text{prod}})^\top$ is the least square estimator, we have

$$\|\mathbf{H} - \widehat{\Theta}\|_F^2 \leq \|\mathbf{H} - \Pi_{\mathcal{T}}(\Theta^*)\|_F^2. \quad (17)$$

Let us define the mean-zero “noise” matrix $\mathbf{E} = \mathbf{H} - \mathbb{E}[\mathbf{H}] = \mathbf{H} - \Theta^*$. We infer from (17) that

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \leq \|\Theta^* - \Pi_{\mathcal{T}}(\Theta^*)\|_F^2 + 2\langle \widehat{\Theta} - \Theta^*, \mathbf{E} \rangle + 2\langle \Theta^* - \Pi_{\mathcal{T}}(\Theta^*), \mathbf{E} \rangle. \quad (18)$$

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 $\langle \widehat{\Theta} - \Theta^*, \mathbf{E} \rangle$. To this end, we define

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

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

$$\langle \widehat{\Theta} - \Theta^*, \mathbf{E} \rangle = \underbrace{\langle \Pi_{\widehat{\mathcal{T}}}(\Theta^*) - \Theta^*, \mathbf{E} \rangle}_{\Xi_1} + \underbrace{\langle \widehat{\Theta} - \Pi_{\widehat{\mathcal{T}}}(\Theta^*), \mathbf{E} \rangle}_{\Xi_2}. \quad (19)$$

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

$$\mathbb{E}(\Xi_1) \leq \sigma \sqrt{2(n \log K + m \log L + 1)} \mathbb{E}[\|\widehat{\Theta} - \Theta^*\|_F^2]^{1/2} + b\rho(n \log K + m \log L + 1).$$

Proof. The main steps of the proof consist in applying the Bernstein inequality to Ξ_1 for a fixed Θ instead of $\Pi_{\widehat{\mathcal{T}}}(\Theta^*)$, using the union bound and then integrating the high-probability bound. For the first step, let $\Theta \in \mathbb{R}^{n \times m}$ satisfy $\Theta_{i,j} \in [0, \rho]$ for every $(i, j) \in [n] \times [m]$. By definition of the inner product, we have $\langle \Theta - \Theta^*, \mathbf{E} \rangle = \sum_{(i,j) \in [n] \times [m]} (\Theta - \Theta^*)_{ij} E_{ij}$. The random variables E_{ij} are independent and satisfy the (σ^2, b) -Bernstein condition. The nm -vector with entries $(\Theta - \Theta^*)_{ij}$ has an infinity norm bounded by ρ . Therefore, the version of the Bernstein inequality stated in (Donier-Meroz et al., 2024, Lemma 12) implies that for all $x > 0$, we have

$$\mathbb{P}(\langle \Theta - \Theta^*, \mathbf{E} \rangle \geq \sqrt{2x} \sigma \|\Theta - \Theta^*\|_F + b\rho x) \leq e^{-x}.$$

Let us define $\Omega_{\mathbf{Z}, \mathbf{Z}'} = \{(\widehat{\mathbf{Z}}^{\text{cust}}, \widehat{\mathbf{Z}}^{\text{prod}}) = (\mathbf{Z}, \mathbf{Z}')\}$, 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 $\Pi_{\widehat{\mathcal{T}}}(\Theta^*)$ is deterministic and its elements are arithmetic means of the

elements of Θ^* . Hence, $0 \leq (\Pi_{\hat{\mathcal{T}}}(\Theta^*))_{ij} \leq \|\Theta^*\|_\infty$ and

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

Note also that the cardinality of $\mathcal{Z}_{n,K,n_0}$ is at most K^n . Therefore, the union bound leads to

$$\begin{aligned} & \mathbb{P}\left(\langle \Pi_{\hat{\mathcal{T}}}(\Theta^*) - \Theta^*, \mathbf{E} \rangle \geq \sqrt{2x\rho} \sigma \|\Pi_{\hat{\mathcal{T}}}(\Theta^*) - \Theta^*\|_F + b\rho\sigma^2 x\right) \\ & \leq \sum_{(\mathbf{Z}, \mathbf{Z}')} \mathbb{P}\left(\left\{\langle \Pi_{\hat{\mathcal{T}}}(\Theta^*) - \Theta^*, \mathbf{E} \rangle \geq \sqrt{2x} \sigma \|\Pi_{\hat{\mathcal{T}}}(\Theta^*) - \Theta^*\|_F + b\rho x\right\} \cap \Omega_{\mathbf{Z}, \mathbf{Z}'}\right) \leq K^n L^m e^{-x}, \end{aligned}$$

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 $uv \leq \lambda u^2 + v^2/(4\lambda)$ entails

$$\mathbb{P}\left(\Xi_1 \geq \lambda \|\Pi_{\hat{\mathcal{T}}}(\Theta^*) - \Theta^*\|_F^2 + \left(\frac{\sigma^2}{2\lambda} + b\rho\right)(n \log K + m \log L + t)\right) \leq e^{-t}$$

for any $\lambda > 0$. Lemma 11 from (Donier-Meroz et al., 2024) yields

$$\begin{aligned} \mathbb{E}(\Xi_1) & \leq \lambda \mathbb{E}\left[\|\Pi_{\hat{\mathcal{T}}}(\Theta^*) - \Theta^*\|_F^2\right] + \left(\frac{\sigma^2}{2\lambda} + b\rho\right)(n \log K + m \log L + 1) \\ & \leq \lambda \mathbb{E}\left[\|\hat{\Theta} - \Theta^*\|_F^2\right] + \left(\frac{\sigma^2}{2\lambda} + b\rho\right)(n \log K + m \log L + 1). \end{aligned}$$

Optimizing with respect to $\lambda > 0$, we get the claim of the lemma. $\square$

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

$$\mathbb{E}(\Xi_2) = \mathbb{E}\left[\langle \Pi_{\hat{\mathcal{T}}}(\mathbf{H}) - \Pi_{\hat{\mathcal{T}}}(\Theta^*), \mathbf{E} \rangle\right] = \mathbb{E}\left[\langle \Pi_{\hat{\mathcal{T}}}(\mathbf{E}), \mathbf{E} \rangle\right] = \mathbb{E}\left[\|\Pi_{\hat{\mathcal{T}}}(\mathbf{E})\|_F^2\right].$$

Lemma 2 (Donier-Meroz et al. (2024)). *Under the conditions of Theorem 1, we have*

$$\mathbb{E}\left[\|\Pi_{\hat{\mathcal{T}}}(\mathbf{E})\|_F^2\right] \leq 4(b + \sigma^2)(3KL + n \log K + m \log L)(2\rho + b\psi_{n,m}(n_0, m_0)).$$

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 (18) and (19). Since $\mathbb{E}[\langle \Theta^* - \Pi_{\mathcal{T}}(\Theta^*), \mathbf{E} \rangle] = \langle \Theta^* - \Pi_{\mathcal{T}}(\Theta^*), \mathbb{E}[\mathbf{E}] \rangle = 0$, we arrive at

$$\begin{aligned} \mathbb{E}[\|\hat{\Theta} - \Theta^*\|_F^2] & \leq \|\Theta^* - \Pi_{\mathcal{T}}\Theta^*\|_F^2 + 2\mathbb{E}[\Xi_1] + 2\mathbb{E}[\Xi_2] \\ & \leq \|\Theta^* - \Pi_{\mathcal{T}}\Theta^*\|_F^2 + 2\sigma\sqrt{2(A+1)}\mathbb{E}[\|\hat{\Theta} - \Theta^*\|_F^2]^{1/2} + 2b\rho(A+1) + 2\mathbb{E}[\Xi_2]. \end{aligned}$$

One can check that the last inequality leads to

$$(\mathbb{E}[\|\hat{\Theta} - \Theta^*\|_F^2]^{1/2} - \sigma\sqrt{2(A+1)})^2 \leq \|\Theta^* - \Pi_{\mathcal{T}}(\Theta^*)\|_F^2 + 2(A+1)(\sigma^2 + b\rho) + 2\mathbb{E}[\Xi_2].$$

This readily yields

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

where in the second line we have used the inequality $\sqrt{x} + \sqrt{y} \leq \sqrt{2x + 2y}$. Finally, under the condition $\psi_{n,m}(n_0, m_0) \leq (\sigma/b)^2$, we get the claim of Theorem 1.

Funding

This work was partially supported by the grant Investissements d’Avenir (ANR-11-IDEX0003/Labex Ecodec/ANR-11-LABX-0047) and funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme grant agreement 741467-FIRMNET.

References

- Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
- Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.
- G rard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473, 2003. Biometrics.
- 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.
- 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 *COLT*, pages 391–423, 2015.
- Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1): 215–237, 2015.
- Jing Lei. A goodness-of-fit test for stochastic block models. *Ann. Statist.*, 44(1):401–424, 2016.
- Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *Ann. Statist.*, 44(5):2252–2280, 2016.
- Y. X. Rachel Wang and Peter J. Bickel. Likelihood-based model selection for stochastic block models. *Ann. Statist.*, 45(2):500–528, 2017.
- Yudong Chen, Xiaodong Li, and Jiaming Xu. Convexified modularity maximization for degree-corrected stochastic block models. *Ann. Statist.*, 46(4):1573–1602, 2018.
- 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.
- Chao Gao, Zongming Ma, Anderson Zhang, and Harrison Zhou. Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(60):1–45, 2017.
- Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.
- 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.
- 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.
- Stefan Neumann. Bipartite stochastic block models with tiny clusters. In *Advances in NeurIPS 2018*, pages 3871–3881, 2018.

- 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, 2019a.
- Zhixin Zhou and Arash A. Amini. Optimal bipartite network clustering. *J. Mach. Learn. Res.*, 21:40:1–40:68, 2020.
- 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.
- 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.
- 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.
- Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation, 2013.
- 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.
- Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6): 2624–2652, 2015.
- 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.
- 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.
- Olga Klopp and Nicolas Verzelen. Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, 174:1033–1090, 2019.
- Chao Gao and Zongming Ma. Minimax Rates in Network Analysis: Graphon Estimation, Community Detection and Hypothesis Testing. *Statistical Science*, 36(1):16 – 33, 2021.
- Yuetian Luo and Chao Gao. Computational lower bounds for graphon estimation via low-degree polynomials. *The Annals of Statistics*, 52(5):2318 – 2348, 2024.
- David J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981.
- Bryan S. Graham. An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4): 1033–1063, 2017. ISSN 0012-9682.
- Bryan S. Graham. Sparse network asymptotics for logistic regression. *Journal of Multivariate Analysis*, October 2020. URL <https://arxiv.org/pdf/2010.04703.pdf>.
- Laurent Davezies, Xavier D’Haultfœuille, and Yannick Guyonvarch. Empirical process results for exchangeable arrays. *The Annals of Statistics*, 49(2):845 – 862, 2021.
- Etienne Donier-Meroz, Arnak S Dalalyan, Francis Kramarz, Philippe Choné, and Xavier D’Haultfœuille. Proofs of the results stated in “Graphon Estimation in bipartite graphs with observable weights and unobservable node labels”. Technical report, November 2024.
- G. Leung and A.R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.
- 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.
- 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.
- 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.
- 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.
- A. S. Dalalyan. Simple proof of the risk bound for denoising by exponential weights for asymmetric noise distributions. *Izv. Nats. Akad. Nauk Armenii Mat.*, 58(6):3–14, 2023.
- S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
- Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.
- Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd’s algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

- Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k -means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.
- Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.
- George B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.
- Narendra Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.
- S. Mehrotra. An implementation of a primal-dual interior point method for linear programming. *ORSA Journal on Computing*, 4(4):470–500, 1992.
- T. S. Motzkin, H. Raiffa, G. L. Thompson, and R. M. Thrall. The double description method. *Contributions to the Theory of Games*, 2:51–73, 1953.
- David Avis and Komei Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polytopes. *Discrete Computational Geometry*, 16(4):391–424, 1996.
- Gurobi Optimizer Reference Manual*. Gurobi Optimization, LLC, Beaverton, OR, USA, 2023.
- IBM ILOG CPLEX Optimization Studio*. IBM Corporation, Armonk, NY, USA, 2023.
- Michael Grant and Stephen Boyd. Disciplined convex programming. *Global Optimization: From Theory to Implementation*, pages 3–62, 2014.
- 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 2019b. URL <https://jmlr.org/papers/volume20/18-170/18-170.pdf>.
- Sivaraman Balakrishnan, Martin J Wainwright, and Ying Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.