Maximum Likelihood Estimation of Sparse Networks with Missing Observations

Solenne Gaucher, Olga Klopp

▶ **To cite this version:**

Solenne Gaucher, Olga Klopp. Maximum Likelihood Estimation of Sparse Networks with Missing Observations. 2019. hal-02050003v1

HAL Id: hal-02050003

https://hal.science/hal-02050003v1

Preprint submitted on 26 Feb 2019 (v1), last revised 27 Apr 2021 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Maximum Likelihood Estimation of Sparse Networks with Missing Observations

Solenne Gaucher [*1] and Olga Klopp [†2,1]

[1]CREST, ENSAE
[2]ESSEC Business School

February 26, 2019

Abstract

Estimating the matrix of connections probabilities is one of the key questions when studying sparse networks. In this work, we consider networks generated under the sparse graphon model and the inhomogeneous random graph model with missing observations. Using the Stochastic Block Model as a parametric proxy, we bound the risk of the maximum likelihood estimator of network connections probabilities, and show that it is minimax optimal. When risk is measured in Frobenius norm, no estimator running in polynomial time has been shown to attain the minimax optimal rate of convergence for this problem. Thus, maximum likelihood estimation is of particular interest as computationally efficient approximations to it have been proposed in the literature and are often used in practice.

Keywords— Missing observations, network models, sparse estimation, graphon model

1 Introduction

In the past two decades, networks have attracted considerable attention, as many scientific fields are concerned by the advances made in the understanding of these complex systems. In social sciences [52] as in physics [3] and biology [54], networks are used to represent a great variety of systems of interactions between social agents, particles, proteins or neurons. These networks are often modeled as an observation drawn from a random graph.

Missing observations is a common problem when studying real life networks. In social sciences, data coming from sample surveys are likely to be incomplete, especially, when dealing with large or hard-to-find populations. While biologists often use graphs to model interactions between proteins, experimental discovery of these interactions can require substantial time and investment from the scientific community [10]. In many cases, collecting complete information on relations between actors can be difficult, expensive and time-consuming [33, 55, 26, 23]. On the other hand, the emergence of detailed data sets coming, for example, from social networks or genome sequencing has fostered new challenges, as their large size makes using the full data computationally unattractive. This has lead scientists to consider only sub-samples of the available data [7]. However, incomplete observation of the network structure may considerably affect the accuracy of inference methods [32].

Our work focuses on the study of the inhomogeneous random graph model with *missing observations*. In this setting, the problem of estimating the matrix of connections probabilities is of primary interest. Minimax optimal convergence rates for this problem have been shown to be attained by the least square estimator under full observation of the network for dense graphs in [20] and for sparse graphs in [28]. In [19], the authors extended these results to the setting in which observations about the presence or absence of an edge are missing independently uniformly at random. Unfortunately, least square estimation is too costly to be used in practice. Many other approaches have been proposed, for example, spectral clustering [41, 24, 47], modularity maximization [44, 9], belief propagation [18], neighborhood smoothing [56], convex relaxation of k-means clustering [22] and of likelihood maximization [5], and universal singular value thresholding [14, 30, 53]. An important question here is the possible computational gap when

[*]solenne.gaucher@ensae.fr
[†]olga.klopp@math.cnrs.fr

no polynomial time algorithm can achieve minimax optimal rate of convergence. The present work is a step further in the understanding of this problem.

In this work, we consider the maximum likelihood estimator. This estimator is also NP-hard but its computationally efficient approximations (under some additional conditions) have been proposed in the literature (see, e.g., [40] for a detailed review of these methods). For example, the authors of [4] suggest to use pseudo-likelihood methods, as it leads to computationally tractable estimators. Alternatively, in [13] a tractable variational approximation of the maximum likelihood estimator is proposed. This methods has been applied successfully to study biological networks, political blogsphere networks and seeds exchange networks [46, 48, 34]. The authors of [8] show asymptotic normality of the maximum likelihood estimate and of its variational approximation for sparse graphs generated by stochastic block models when the connections probabilities of the different communities are well separated. In [48], these results are extended to the case of missing observations. These methods suffer from a lack of theoretical guarantees when the model is misspecified or non-identifiable. On the other hand, to the best of our knowledge, no non-asymptotic bound has been established for the risk of the maximum likelihood estimator. In this work, we close this gap and show that the maximum likelihood estimator is minimax optimal in a number of scenarii.

Our results also find a natural application in predicting the existence of non-observed edges, a commonly encountered problem called *link prediction* [38, 57]. Interaction networks are often incomplete, as detecting interactions can require significant experimental effort. Instead of exhaustively testing for every connection, one might be interested in deducing the pairs of agents which are most likely to interact based on the relations already recorded and on available covariates. If these estimations are precise enough, testing for these interactions would enable scientists to establish the network topology while substantially reducing the costs [16]. In this context, estimating the probabilities of connections through likelihood maximization enables to accordingly rank unobserved pairs of nodes. Link prediction also finds applications in recommender systems for social networks [50]. The missing observation scheme studied in this work is motivated by the above examples, and generalizes the model described in [19].

1.1 Inhomogeneous random graph model

We consider an undirected, unweighted graph with n nodes indexed from 1 to n. Its connectivity can be encoded by its *adjacency matrix* $\boldsymbol{A}$, defined as follows: set $\boldsymbol{A}$ a $n \times n$ symmetric matrix such that for all $i < j$, $\boldsymbol{A}_{ij} = 1$ if there exists an edge between node i and node j, $\boldsymbol{A}_{ij} = 0$ otherwise. In our model, we consider that there is no edge linking a node to itself, so $\boldsymbol{A}_{ii} = 0$ for all i. We assume that the variables $(\boldsymbol{A}_{ij})_{1 \leq i < j \leq n}$ are independent Bernoulli random variables of parameter $\boldsymbol{\Theta}^*_{ij}$, where $\boldsymbol{\Theta}^*$ is a $n \times n$ symmetric matrix with zero diagonal entries. This matrix $\boldsymbol{\Theta}^*$ corresponds to the matrix of probabilities of observing an edge between nodes i and j. This model is known as the *inhomogeneous random graph* model:

$$\forall 1 \leq i < j \leq n, \ \boldsymbol{A}_{ij} | \boldsymbol{\Theta}^*_{ij} \overset{ind.}{\sim} \text{Bernoulli}\left(\boldsymbol{\Theta}^*_{ij}\right). \tag{1}$$

In the present paper we consider the following problem: from a single partial observation of the graph, that is, given a sample of entries of the adjacency matrix $\boldsymbol{A}$, we want to estimate the matrix of connections probabilities $\boldsymbol{\Theta}^*$.

The problem of estimating $\boldsymbol{\Theta}^*$ when some entries of the adjacency matrix are not observed is closely related to the 1-Bit matrix completion problem. The matrix completion problem [12, 31, 43] aims at recovering a matrix which is only partially observed. More precisely, we observe a random sample of its entries, which may be corrupted by some noise, and we wish to infer the rest of the matrix. In 1-bit matrix completion, first introduced in [17], the entries (i, j) of the observed matrix can only take two values $\{0, 1\}$ with probabilities given respectively by $f(\boldsymbol{M}_{ij})$ and $1 - f(\boldsymbol{M}_{ij})$. Here, the matrix $\boldsymbol{M}$ corresponds to the real quantity of interest that one would like to infer, and the function f can be seen as the cumulative distribution function of the noise. A typical assumption in this setting is that the matrix $\boldsymbol{M}$ is low-rank. In [29], the authors show that for 1-bit matrix completion the restricted penalized maximum likelihood estimator is minimax optimal up to a log factor. The methods used in our proofs are, to some extend, inspired by the methods developed for the framework of matrix completion. However, the problem we have in hand is in many aspects different from the 1-bit matrix completion problem. The structure of the connections probabilities matrix $\boldsymbol{\Theta}^*$ and the sparsity of the network allow for faster rates of convergence, and the technics of proof required to match the minimax optimal convergence rate are more involved.

Our approach for estimating the matrix of connections probabilities is based on the celebrated Regularity Lemma by Szemerédi [37], which implies that any graph can be well approximated by a stochastic block model (SBM). We refer to [37] for a more detailed presentation of this result. In the SBM, each node i is associated with a community $z^*(i)$, where $z^* : [n] \to [k]$ is called the index function. This index function can either be treated as a parameter to estimate (this model is sometimes called the conditional stochastic block model), or as a latent variable. In this case, the indexes follow a multinomial distribution: $\forall i, \ z^*(i) \overset{i.i.d}{\sim} \mathcal{M}(1; \alpha^*)$ where $\forall l \in [k]$, α_l is the probability that node i belongs to the community l. Given this index function, the probability that there exists an edge between nodes i and j depends only on the communities of i and j. For example, when considering citations networks, where two

articles are linked if one is cited by the other, it amounts to saying that the probability that two articles are linked only depends on their topic. Similarly, if one considers students of a school in a social network, it is a reasonable assumption to say that the probability that two students are linked only depends on their cohorts. This implies that the matrix of connections probabilities $\boldsymbol{\Theta}^*$ can be factorized as follows: $\boldsymbol{\Theta}^*_{ij} = \boldsymbol{Q}^*_{z^*(i)z^*(j)}$, with $\boldsymbol{Q}^*$ a $k \times k$ symmetric matrix such that $\boldsymbol{Q}^*_{ab}$ is the probability that there exists an edge between a given member of the community a and a given member of the community b, so we have that the conditional SBM can be written as:

$$\begin{gathered} \exists \boldsymbol{Q}^* \in [0,1]^{k\times k}_{\text{sym}},\ \exists z^* : [n] \to [k] \\ \forall 1 \leq i < j \leq n,\ \boldsymbol{A}_{ij} | \boldsymbol{Q}^*, z^* \overset{ind.}{\sim} \text{Bernoulli}\left(\boldsymbol{Q}^*_{z^*(i)z^*(j)}\right),\ \boldsymbol{A}_{ii} = 0. \end{gathered} \tag{2}$$

While considering the SBM, the problem of estimating the matrix of connections probabilities reduces to estimating the label function z^* and the matrix of probabilities of connections between communities $\boldsymbol{Q}^*$.

In the past decade, the stochastic block model has known a growing interest from the statistical community and an important part of the work has focused on the problem of community recovery (i.e., the recovery of the vector of communities populations α^*, or of the label function z^* in the conditional model). Theoretical guarantees for this problem were established under quite strong assumptions on the matrix of probabilities of connections between communities, $\boldsymbol{Q}^*$, see, for example, [39, 11, 1, 42].

Note that our results hold without assuming the existence of the true community structure, that is, without assuming that the matrix $\boldsymbol{\Theta}^*$ is block constant. With this in mind, we will focus on estimating the distribution giving rise to the adjacency matrix, i.e., on estimating $\boldsymbol{\Theta}^*$, rather than on estimating the label function or the populations of the communities. One important question in this setting is how to choose the number of communities for our estimator, as more communities implies a smaller bias and a greater variance. Optimizing this trade-off requires, first, establishing a non-asymptotic bound on the risk of our estimator for a number of communities that may depend on the number of nodes, and, in a second time, bounding the bias of an oracle block constant estimator.

Our work focuses on relevant in applications setting of partial observations of the network. We consider the following missing value setting. Let $\boldsymbol{X} \in \{0,1\}^{n\times n}_{sym}$ denote the sampling matrix given by $\boldsymbol{X}_{ij} = 1$ if we observe $\boldsymbol{A}_{ij}$ and $\boldsymbol{X}_{ij} = 0$ otherwise. We assume that the sampling matrix $\boldsymbol{X}$ is random and, conditionally on $\boldsymbol{\Theta}^*$, independent from the adjacency matrix $\boldsymbol{A}$. For all $1 \leq i < j \leq n$, its entries $\boldsymbol{X}_{ij}$ are mutually independent. Finally, we denote by $\boldsymbol{\Pi} \in [0,1]^{n\times n}_{sym}$ the matrix of sampling probabilities such that $\boldsymbol{X}_{ij} \overset{ind.}{\sim} \text{Bernoulli}(\boldsymbol{\Pi}_{ij})$. This sampling scheme includes for instance node-based sampling schemes such as the exo-centered design described in [26], where we observe $\boldsymbol{A}_{ij}$ if i or j belongs to the set of sampled nodes. It also covers random dyad sampling schemes (described, e.g., in [48]). In this case, the probability of observing the entry $\boldsymbol{A}_{ij}$ is allowed to depend on the communities of i and j.

1.2 Graphon model

While studying exchangeable random graphs, important questions such as how to compare two graphs with different numbers of nodes or how to study graphs with an increasing number of nodes call for a more general, non-parametric model. One of such models that has attracted a lot of attention recently is the *graphon* model [45, 20, 28, 53]. In this model, the connections probabilities $\boldsymbol{\Theta}^*_{ij}$ are the following random variables

$$\boldsymbol{\Theta}^*_{ij} = W^*(\zeta_i, \zeta_j) \tag{3}$$

where $\zeta_1, ..., \zeta_n$ are unobserved (latent) independent random variables sampled uniformly in $[0,1]$. The graph is then sampled according to the inhomogeneous random graph model (1). The function $W^* : [0,1]^2 \to [0,1]$ is measurable, symmetric and is called a graphon. Graphs encountered in practice are usually *sparse*: the expected number of edges grows as $\rho_n n^2$ where ρ_n is a decreasing sequence of sparsity inducing parameters. The dense graphon model can be modified in order to account for this sparsity:

$$\boldsymbol{\Theta}^*_{ij} = \rho_n W^*(\zeta_i, \zeta_j). \tag{4}$$

Since the law of the graph is invariant under any change of labelling of its nodes, different graphons can give rise to the same distribution on the space of graphs of size n. More precisely, let W be a graphon and $\tau : [0,1] \to [0,1]$ be a measure-preserving function. We write $W_\tau(x,y) = W(\tau(x), \tau(y))$ and say that two graphons U and V are *weakly isomorphic* if there exists measure-preserving maps τ , ϕ such that $U_\tau = W_\phi$ almost everywhere. It is established in Section 10, [37] that two graphons define the same probability measure on graphs if and only if they are weakly isomorphic.

In the present paper we also consider the setting when the matrix of connections probabilities is generated following the sparse graphon model (4). We deal with two classes of graphon functions previously studied in the literature, step-function graphons and smooth graphons, under the scenario of partial observations of the network.

1.3 Outline of the paper

The present paper is devoted to the theoretical study of the maximum likelihood estimator in sparse network models with missing observations. First, we provide an oracle bound for the risk of the maximum likelihood estimator of the matrix of connections probabilities from a partial observation of the adjacency matrix $\boldsymbol{A}$. Our results hold under fairly general assumptions on the missing observations scheme and we show that the maximum likelihood estimator matches the minimax optimal rates of convergence in a variety of scenarii. Second, we provide an adaptive version of our estimator which, in particular, does not require the knowledge of the sparsity parameter ρ_n. We also bound the Kullback-Leibler divergence between the true matrix of connections probabilities and its block constant approximation, and derive an optimal choice for the number of communities defining the maximum likelihood estimator.

This manuscript is organized as follows. In Section 2.1, we introduce the maximum likelihood estimator for the matrix of connections probabilities $\boldsymbol{\Theta}^*$ from partial observation of the adjacency matrix $\boldsymbol{A}$. Then, Theorem 1 in Section 2.2 provides a non-asymptotic oracle bound on the risk of this estimator. As a consequence, we show that our estimator is minimax optimal in a number of scenarii and derive the corresponding bound for estimating $\boldsymbol{\Theta}^*$ in the case of full observation of the adjacency matrix $\boldsymbol{A}$. Our estimation method requires bounds on the entries of $\boldsymbol{\Theta}^*$. In Section 2.3, we first propose a method to choose these bounds under fairly general assumptions and, in Section 2.4, we specify it to the case of sparse graphon model (4). We show that the resulting adaptative estimator is minimax optimal up to a log factor. Finally, in Section 2.5, Theorem 4, we provide the choice for the number of communities that achieves the best trade off between the variability of our estimate and the fit of the oracle model.

1.4 Notations

We provide here a summary of the notations used throughout this paper.work.

- For any positive integer d, we denote by $[d]$ the set $\{1, ..., d\}$.
- For any set $\mathcal{S}$, we denote by $|\mathcal{S}|$ its cardinality.
- For any matrix $\boldsymbol{A}$, we denote by $\boldsymbol{A}_{ij}$ its entry on row i and column j. If $\boldsymbol{A} \in [0,1]^{n\times n}$ and $\boldsymbol{A}$ is symmetric, we write $\boldsymbol{A} \in [0,1]^{n\times n}_{\text{sym}}$.
- Let $\mathcal{K}(q, q') = q \log\left(\frac{q}{q'}\right) + (1-q)\log\left(\frac{1-q}{1-q'}\right)$ denote the Kullback-Leibler divergence of a Bernoulli distribution with parameter q from a Bernoulli distribution with parameter q'. For any three symmetric matrices with zero diagonal entries $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{X} \in [0,1]^{n\times n}_{sym}$ we set
$$\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i<j} \boldsymbol{X}_{ij}\mathcal{K}(\boldsymbol{A}_{ij}, \boldsymbol{B}_{ij}) \quad \text{and} \quad \mathcal{K}(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i<j} \mathcal{K}(\boldsymbol{A}_{ij}, \boldsymbol{B}_{ij}).$$
- For any three symmetric matrices with zero diagonal entries $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{X} \in \mathbb{R}[0,1]^{n\times n}_{sym}$, let $\langle \boldsymbol{A}|\boldsymbol{B}\rangle = \sum_{i<j} \boldsymbol{A}_{ij}\boldsymbol{B}_{ij}$, $\langle \boldsymbol{A}|\boldsymbol{B}\rangle_{\boldsymbol{X}} = \sum_{i<j} \boldsymbol{X}_{ij}\boldsymbol{A}_{ij}\boldsymbol{B}_{ij}$, $\|\boldsymbol{A}\|_2 = \sqrt{\langle \boldsymbol{A}|\boldsymbol{A}\rangle}$, $\|\boldsymbol{A}\|_{2,\boldsymbol{X}} = \sqrt{\langle \boldsymbol{A}|\boldsymbol{A}\rangle_{\boldsymbol{X}}}$, and $\|\boldsymbol{A}\|_\infty = \max_{i,j}|\boldsymbol{A}_{ij}|$.
- We denote by $\mathcal{Z}_{n,k}$ the label functions $z : [n] \to [k]$. For any $z \in \mathcal{Z}_{n,k}$, we denote by $\mathcal{T}_z$ the set of block constant matrices corresponding to the label z:
$\mathcal{T}_z \triangleq \left\{ \boldsymbol{A} : \forall i \in [n], \boldsymbol{A}_{ii} = 0 \ \& \ \exists \boldsymbol{Q} \in [0,1]^{k\times k}, \forall 1 \le i < j \le n, \boldsymbol{A}_{ij} = \boldsymbol{A}_{ji} = \boldsymbol{Q}_{z(i)z(j)} \right\}$.
- To ease notations, for $\boldsymbol{A} \in \mathcal{T}_z$ and $(a,b) \in [k]^2$, we sometimes denote by $\boldsymbol{A}_{z^{-1}(a)z^{-1}(b)}$ any entry $\boldsymbol{A}_{ij}$ such that $(i,j) \in (z^{-1}(a), z^{-1}(b))$ and $i \neq j$. We write $\mathcal{T}_k = \bigcup_{z\in\mathcal{Z}_{n,k}} \mathcal{T}_z$.
- We denote by C and C' positive constants that can vary from line to line. These are absolute constants unless otherwise mentioned.
- We denote respectively by $\mathbb{E}^{\boldsymbol{X}}$ and $\mathbb{P}^{\boldsymbol{X}}$ the expectation and the probability conditionally on the random variable $\boldsymbol{X}$, and respectively by $\mathbb{E}$ and $\mathbb{P}$ the expectation and the probability over all random variables.

2 Convergence rate for the maximum likelihood estimator

2.1 Maximum likelihood estimator under missing observations

We start by introducing the conditional log-likelihood for the model (1). Conditionally on the probability matrix $\boldsymbol{\Theta}^*$, the entries $(\boldsymbol{A}_{ij})_{1\le i<j\le n}$ of the adjacency matrix $\boldsymbol{A}$ are independent Bernoulli variables with parameters

$(\boldsymbol{\Theta}^*_{ij})_{1\leq i<j\leq n}$. Therefore, for any $\boldsymbol{\Theta}\in[0,1]^{n\times n}$, the conditional log-likelihood of the parameter matrix $\boldsymbol{\Theta}$ with respect to the observed entries of the adjacency matrix $\boldsymbol{A}$ is given by

$$\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};\boldsymbol{\Theta})=\sum_{i<j}\boldsymbol{X}_{ij}\left(\boldsymbol{A}_{ij}\log(\boldsymbol{\Theta}_{ij})+(1-\boldsymbol{A}_{ij})\log(1-\boldsymbol{\Theta}_{ij})\right).$$

For any $z\in\mathcal{Z}_{n,k}$ and $Q\in[0,1]^{k\times k}_{\text{sym}}$, the matrix of connections probabilities corresponding to the block model $(z,\boldsymbol{Q})$ is given by $\boldsymbol{\Theta}_{ij}=\boldsymbol{Q}_{z(i)z(j)}$ for $1\leq i<j\leq n$ and $\boldsymbol{\Theta}_{ii}=0$ for $i\in[n]$. With these notations, the conditional log-likelihood of a block model $(z,\boldsymbol{Q})$ with respect to the observed entries of the adjacency matrix $\boldsymbol{A}$ is

$$\begin{aligned}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};z,\boldsymbol{Q})&=\sum_{1\leq i<j\leq n}\boldsymbol{X}_{ij}\left(\boldsymbol{A}_{ij}\log(\boldsymbol{Q}_{z(i)z(j)})+(1-\boldsymbol{A}_{ij})\log(1-\boldsymbol{Q}_{z(i)z(j)})\right)\\&=\sum_{1\leq a\leq b\leq k}\sum_{\substack{i\in z^{-1}(a),\ j\in z^{-1}(b)\\ i\neq j}}\boldsymbol{X}_{ij}\left(\boldsymbol{A}_{ij}\log(\boldsymbol{Q}_{ab})+(1-\boldsymbol{A}_{ij})\log(1-\boldsymbol{Q}_{ab})\right)\\&=\sum_{a\leq b}\log(\boldsymbol{Q}_{ab})\sum_{\substack{i\in z^{-1}(a),\ j\in z^{-1}(b)\\ i\neq j}}\boldsymbol{X}_{ij}\boldsymbol{A}_{ij}+\sum_{a\leq b}\log(1-\boldsymbol{Q}_{ab})\sum_{\substack{i\in z^{-1}(a),\ j\in z^{-1}(b)\\ i\neq j}}\boldsymbol{X}_{ij}(1-\boldsymbol{A}_{ij}).\end{aligned}$$

The maximum likelihood estimator for the stochastic block model is

$$(\widehat{\boldsymbol{Q}},\widehat{z})\in\operatorname*{arg\,max}_{\boldsymbol{Q}\in[0,1]^{k\times k}_{\text{sym}},z\in\mathcal{Z}_{n,k}}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};z,\boldsymbol{Q}).$$

The block constant maximum likelihood estimator of $\boldsymbol{\Theta}^*$ is defined as $\widehat{\boldsymbol{\Theta}}_{ij}=\widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)}$ for all $i<j$. Note that maximizing the log-likelihood is equivalent to minimizing a sum of Bernoulli Kullback-Leibler divergences. Indeed, an easy calculation leads

$$(\widehat{\boldsymbol{Q}},\widehat{z})\in\operatorname*{arg\,max}_{\boldsymbol{Q}\in[0,1]^{k\times k}_{\text{sym}},z\in\mathcal{Z}_{n,k}}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};z,\boldsymbol{Q})=\operatorname*{arg\,min}_{\boldsymbol{Q}\in[0,1]^{k\times k}_{sym},z\in\mathcal{Z}_{n,k}}\sum_{i<j}\boldsymbol{X}_{ij}\mathcal{K}(\boldsymbol{A}_{ij},\boldsymbol{Q}_{z(i)z(j)}).\tag{5}$$

Moreover, for any fixed assignment $z\in\mathcal{Z}_{n,k}$ and any sampling matrix $\boldsymbol{X}$, the log-likelihood with regards to the observed entries of $\boldsymbol{A}$ will be maximized by taking $\boldsymbol{Q}_{ab}=\overline{\boldsymbol{X}\boldsymbol{A}}^z_{ab}\triangleq\frac{\sum_{i\in z^{-1}(a),j\in z^{-1}(b)\ i\neq j}\boldsymbol{X}_{ij}\boldsymbol{A}_{ij}}{\sum_{i\in z^{-1}(a),j\in z^{-1}(b)\ i\neq j}\boldsymbol{X}_{ij}}$:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};z)&=\max_{\boldsymbol{Q}\in[0,1]^{k\times k}_{\text{sym}}}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};z,\boldsymbol{Q})\\&=\sum_{a\leq b}\left(\sum_{\substack{i\in z^{-1}(a),j\in z^{-1}(b)\\ i\neq j}}\boldsymbol{X}_{ij}\right)\left(\overline{\boldsymbol{X}\boldsymbol{A}}^z_{ab}\log(\overline{\boldsymbol{X}\boldsymbol{A}}^z_{ab})+(1-\overline{\boldsymbol{X}\boldsymbol{A}}^z_{ab})\log(1-\overline{\boldsymbol{X}\boldsymbol{A}}^z_{ab})\right).\end{aligned}$$

Under full observation of the network, that is, when for all $1\leq i<j\leq n$, $\boldsymbol{X}_{ij}=1$, previous work [20, 28] on minimax estimation of the matrix of connections probabilities considered the least square estimator

$$(\widehat{\boldsymbol{Q}},\widehat{z})\in\operatorname*{arg\,min}_{\boldsymbol{Q}\in[0,1]^{k\times k}_{\text{sym}},z\in\mathcal{Z}_{n,k}}\sum_{1\leq i<j\leq n}\left(\boldsymbol{A}_{ij}-\boldsymbol{Q}_{z(i)z(j)}\right)^2.$$

Note that for any label function z, the least square criterion will be minimized by taking $\boldsymbol{Q}_{ab}=\overline{\boldsymbol{X}\boldsymbol{A}}^z_{ab}$. Thus, the possible difference between the log-likelihood estimator and the least square estimator lies in the label function that they select.

In the rest of this work, we will denote by $\widetilde{\boldsymbol{\Theta}}$ the oracle probability matrix, i.e., the best approximation to $\boldsymbol{\Theta}^*$ in the sense of the weighted Kullback Leibler divergence:

$$\begin{aligned}&\widetilde{\boldsymbol{\Theta}}_{i<j}=\boldsymbol{Q}^*_{z^*(i)z^*(j)},\widetilde{\boldsymbol{\Theta}}_{ii}=0\\&(\boldsymbol{Q}^*,z^*)\in\operatorname*{arg\,min}_{\boldsymbol{Q}\in[0,1]^{k\times k}_{\text{sym}},z\in\mathcal{Z}_{n,k}}\sum_{i<j}\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*_{ij},\boldsymbol{Q}_{z(i)z(j)}).\end{aligned}\tag{6}$$

2.2 Upper bound on the risk of the restricted maximum likelihood estimator

In this section, we establish an upper bound on the risk of the maximum likelihood estimator and show that it matches the minimax convergence rate obtained in [28, 19]. We will measure the risk of our estimator in Frobenius norm. To bound the risk of the maximum likelihood estimator, we assume that there exists sequences ρ_n and γ_n such that $\forall i < j$,

$$0 < \gamma_n \leq \mathbf{\Theta}^*_{ij} \leq \rho_n < 1. \tag{7}$$

Note that for sparse graphs, ρ_n corresponds to the sparsity inducing sequence in equation (4). We need condition (7) to have the equivalence between the Frobenius distance and the Kullback-Leibler divergence. This assumption is systematic in the literature studying the maximum likelihood estimator for the stochastic block model as it guarantees that the loss associated to the maximum likelihood estimator is Lipschitz. See, for example, [8] and [51], where the authors assume that the adjacency matrix is generated by an homogeneous stochastic block model for which the matrix $\boldsymbol{Q}^*/\rho_n$ has entries bounded away from 0. In our model, this corresponds to imposing that $\rho_n = O(\gamma_n)$, i.e., that all entries of the matrix of connections probabilities $\mathbf{\Theta}^*$ are of the same order of magnitude. Our assumptions are more general than the one developped in these articles, as they also cover the case $\gamma_n = o(\rho_n)$.

In [13], the authors consider the dense SBM and assume that the entries of $\boldsymbol{Q}^*$ belong to $\{0\} \cup [\zeta, 1-\zeta] \cup \{1\}$ for some $\zeta > 0$. They prove the consistency of the maximum likelihood estimator constrained to a restricted subset of the parameters. However, the definition of this subset implies knowing the set $\Omega_0 = \{(i,j) : \mathbf{\Theta}^*_{ij} \in \{0,1\}\}$ prior to estimating the matrix of connections probabilities. Note that, if we assume that Ω_0 is known and that $\boldsymbol{Q}^*$ belong to $\{0\} \cup [\zeta, 1-\zeta] \cup \{1\}$, we can set $\widehat{\mathbf{\Theta}}_{ij} = 0$ for any $(i,j) \in \Omega_0$ and estimate the remaining entries (which are bounded away from 0 and 1) with our procedure.

On the other hand, cases where the entries of $\mathbf{\Theta}^*$ are of different order of magnitude are common in the literature in the case of planted partition models and assortative and disassortative SBM. In the planted partition model, the matrix of connections probabilities between communities is given by $\boldsymbol{Q}^* = (p-q)\mathbf{I}_k + q\mathbf{1}_k\mathbf{1}_k^T$, where $p > q$, $\mathbf{I}_k$ is the identity matrix and $\mathbf{1}_k\mathbf{1}_k^T$ the matrix whose entries are all equal to 1. This amounts to saying that the probability that two nodes are connected only depends on whether they belong to the same community or not. This model can be relaxed to give rise to the assortative model, where the within group probabilities of connection $\boldsymbol{Q}^*_{aa}$ are larger than the between group probabilities of connection $\boldsymbol{Q}^*_{bc}$: there exists $p, q \in [0,1]$ such that for any $a \neq b$, one has $\boldsymbol{Q}^*_{ab} \leq q < p \leq \boldsymbol{Q}^*_{aa}$. The disassortative model corresponds to the case where between communities connections are more likely than within community connections: one has for any $a \neq b$, $\boldsymbol{Q}^*_{aa} \leq q < p \leq \boldsymbol{Q}^*_{ab}$. The last two models are closely related. Indeed, if $\boldsymbol{A}$ is drawn from an assortative SBM, $\mathbf{1}_n\mathbf{1}_n^T - \mathbf{I}_n - \boldsymbol{A}$ corresponds to a realization of a disassortative SBM.

In the planted partition model, maximizing the likelihood is equivalent to finding a partition maximizing the within group connectivity, i.e., maximizing $\sum_{i<j} \boldsymbol{A}_{ij}\boldsymbol{Z}_{ij}$ where $\boldsymbol{Z}_{ij} = \mathbb{1}\{z(i) = z(j)\}$. Convex relaxations of the constraints on $\boldsymbol{Z}$ have been studied in the literature [25, 6, 2], and theoretical guarantees for these algorithms for the problem of communities recovery have been established under assumptions on the gap $p - q$. In these models, communities are characterized by higher (respectively lower) connectivity, and the assumption that $q \ll p$ actually makes the recovery problem easier. By contrast, the definition of a community in the SBM as a set of nodes with the same stochastic behaviour is far more general. It covers settings not suitably described by assortative or disassortative models, as, e.g., graphs with leaders and followers such as the well known example of Zachary's Karate Club (see, e.g., [35]). In these models, leaders are seldomly linked one to another, but are highly connected to their own set of followers. On the other hand, followers rarely connect one to another or to more than one leader. By comparison, our results hold without any assumption on the assortativity or the disassortativity of the model.

In a first time, we assume that we know γ_n and ρ_n. We will discuss how to estimate these values in Section 2.4. Let $\widehat{\mathbf{\Theta}}$ be the block constant estimator based on the maximization of the likelihood among block constant matrices with entries in $[\gamma_n, \rho_n]$:

$$\widehat{\mathbf{\Theta}}_{i<j} = \widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)}, \ \widehat{\mathbf{\Theta}}_{ii} = 0$$
$$(\widehat{\boldsymbol{Q}}, \widehat{z}) \in \underset{\boldsymbol{Q} \in [\gamma_n, \rho_n]^{k\times k}_{\mathrm{sym}}, z \in \mathcal{Z}_{n,k}}{\arg\min} \sum_{i<j} \boldsymbol{X}_{ij}\mathcal{K}(\boldsymbol{A}_{ij}, \boldsymbol{Q}_{z(i)z(j)}).$$

Here we assume that k is fixed and that it can depend on the number of nodes n. k can be chosen using a network cross-validation method [15] or, when the graphon is a step function, it can be chosen using a sequential goodness-of-fit testing procedure [36] or a likelihood-based model selection method [51]. When graphon is Hölder-continuous, we provide a choice of k to optimize the usual trade-off between bias and variance of our estimator in Section 2.4.

Theorem 1. *Assume that $\boldsymbol{A}$ is drawn according to (1), and that $\rho_n = \omega(n^{-1})$. Then, there exists absolute constants $C, C' > 0$ such that with probability at least $1 - 9\exp\left(-C\rho_n n \log(k)\right)$*

$$\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|^2_{2,\mathbf{\Pi}} \leq C'\rho_n \left(\mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}) + \frac{\rho_n^2}{(1-\rho_n)^2 \wedge \gamma_n^2} \left(k^2 + n\log(k)\right) \right).$$

Remark 1. *Note that we are not interested in regimes for which $\rho_n = O(n^{-1})$, as Theorem 2 implies that in this setting the constant estimator with all entries equal to the average node degree attains the minimax rate.*

Remark 2. *This bound is stated as a function of the of weighted Kullback-Leibler divergence $\mathcal{K}_{\mathbf{\Pi}}$ and of the oracle matrix $\widetilde{\mathbf{\Theta}}$ defined in (6). Note that it implies the weaker bound*

$$\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_{2,\mathbf{\Pi}}^2 \leq C' \rho_n \left(\mathcal{K}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}^f) + \frac{\rho_n^2}{(1-\rho_n)^2 \wedge \gamma_n^2} \left(k^2 + n \log(k)\right) \right)$$

where $\widetilde{\mathbf{\Theta}}^f$ is the oracle matrix for the full Kullback-Leibler divergence $\mathcal{K}$. Indeed, one has $\mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}) \leq \mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}^f) \leq \mathcal{K}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}^f)$.

In the case where all entries are observed, that is $\mathbf{\Pi}_{ij} = 1$ for all $i < j$, the rate attained by the maximum likelihood estimator is given by the following corollary.

Corollary 1. *Assume that $\boldsymbol{A}$ is drawn according to (1), that $\forall 1 \leq i < j \leq n$, $\mathbf{\Pi}_{ij} = 1$ and that $\rho_n = \omega(n^{-1})$. Then, there exists positive constants $C, C' > 0$ such that with probability at least $1 - 9\exp\left(-C\rho_n n \log(k)\right)$*

$$\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_2^2 \leq C' \rho_n \left(\mathcal{K}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}) + \frac{\rho_n^2}{(1-\rho_n)^2 \wedge \gamma_n^2} \left(k^2 + n \log(k)\right) \right).$$

If we assume that the probability of observing any entry of the adjacency matrix is bounded away from 0, Theorem 1 can be adapted to provide a bound on the risk of our estimator under the Frobenius norm. Indeed, if $\min_{1 \leq i < j \leq n} \{\mathbf{\Pi}_{ij}\} \geq p$, then $\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_2^2 \leq \frac{1}{p} \|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_{2,\mathbf{\Pi}}^2$ and we get the following result.

Corollary 2. *Assume that $\boldsymbol{A}$ is drawn according to (1), that $\min_{1 \leq i < j \leq n} \{\mathbf{\Pi}_{ij}\} \geq p$ and that $\rho_n = \omega(n^{-1})$. Then, there exists absolute constants $C, C' > 0$ such that with probability at least $1 - 9\exp\left(-C\rho_n \left(k^2 + n\log(k)\right)\right)$*

$$\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_2^2 \leq C' \frac{\rho_n}{p} \left(\mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}) + \frac{\rho_n^2}{((1-\rho_n)^2 \wedge \gamma_n^2)} \left(k^2 + n \log(k)\right) \right).$$

Previously, the problem of estimation of connections probabilities matrix $\mathbf{\Theta}^*$ from partial observations of the network was studied, in particular, by Gao et al. [19]. In this paper, the authors assume that any entry of the adjacency matrix $\boldsymbol{A}$ is observed independently from the others with the same probability p, which is assumed to be known. They establish the following lower bound on the risk of any estimator for the stochastic block model.

Theorem 2 (Gao et al., 2017)**.** *Assume that $\boldsymbol{A}$ is drawn according to (2), and that each edge is observed independently from the others with probability p. There exists universal constants $C, C' > 0$ such that*

$$\inf_{\widehat{\mathbf{\Theta}}} \sup_{\mathbf{\Theta}^* \in \mathcal{T}_k,\ \|\mathbf{\Theta}^*\|_\infty \leq \rho_n} \mathbb{P}\left[\left\| \mathbf{\Theta}^* - \widehat{\mathbf{\Theta}} \right\|_2^2 \geq C \left(\frac{\rho_n (n \log(k) + k^2)}{p} \wedge \rho_n^2 n^2 \right) \right] > C'.$$

The authors of [19] also prove that the least square estimator is minimax optimal in this setting. Note that our missing data scheme is more general and more realistic than the one studied in [19], and that our estimator does not require information on the probability of observing the entries of the adjacency matrix $\boldsymbol{A}$. In the particular case when $\boldsymbol{X}_{ij} \sim \text{Bernoulli}(p)$ and $\rho_n = O(\gamma_n)$, Corollary 2 and the lower bound in Theorem 2 ensures that the maximum likelihood estimator is minimax optimal. We underline that although the lower bound has been established in [19] for $\mathbf{\Theta} \in \mathcal{T}_k$, $\|\mathbf{\Theta}\|_\infty \leq \rho_n$, its proof can be adapted to provide a lower bound on the convergence rate for a smaller set of parameters $\left\{ \mathbf{\Theta} \in \mathcal{T}_k,\ \|\mathbf{\Theta}\|_\infty \leq \rho_n,\ \min_{i<j} \{\mathbf{\Theta}_{ij}\} \geq \gamma_n \right\}$. Indeed, the "non parametric" as well as the "clustering" components of the rate are established using matrices with entries close to $\frac{\rho_n}{2}$.

2.3 Choice of γ_n under general assumptions

In this section, we deal with the setting when condition $\min_{i<j} \mathbf{\Theta}^*_{ij} > \gamma_n$ is violated. In what follows we consider the sparse case, that is $\rho_n \to 0$, so $\gamma_n \leq 1 - \rho_n$ for n large enough. As discussed in [28], we can easily estimate ρ_n (see also Section 2.4). On the other hand, when some entries of the matrix of connections probabilities $\mathbf{\Theta}^*$ can be 0 or arbitrarily close to 0, choosing the best sequence γ_n comes down to a trade-off between errors caused by estimating entries smaller than γ_n by γ_n, and the bound obtained in Theorem 1. We first consider the case when there exists a sequence γ_n such that number of small entries $n_s = \sum_{i<j} \mathbb{1}\{\widetilde{\mathbf{\Theta}}_{ij} < \gamma_n\}$ is small enough. Then, we have the following result:

Corollary 3. *Assume that $\boldsymbol{A}$ is drawn according to (1), that $\rho_n = \omega(n^{-1})$ and that $n_s \leq \frac{k^2 \vee (n\log(k))}{\rho_n}$. Then, there exists absolute constants $C, C' > 0$ such that with probability at least $1 - 9\exp\left(-C\rho_n\left(k^2 + n\log(k)\right)\right)$*

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 \leq C'\rho_n\left(\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \frac{\rho_n^2}{\gamma_n^2}\left(k^2 + n\log(k)\right)\right).$$

To see it, we define

$$\begin{aligned} \widetilde{\boldsymbol{\Theta}}^s_{ij} &= \boldsymbol{Q}^s_{z^*(i)z^*(j)}, \widetilde{\boldsymbol{\Theta}}^s_{ii} = 0 \\ \boldsymbol{Q}^s_{ab} &= \boldsymbol{Q}^*_{ab} \vee \gamma_n \end{aligned} \tag{8}$$

where $\boldsymbol{Q}^*$ is given by (6). Note that $\widetilde{\boldsymbol{\Theta}}^s$ and $\widehat{\boldsymbol{\Theta}}$ are defined on the same set, and thus $\mathcal{K}_{\boldsymbol{X}}(\boldsymbol{A}, \widehat{\boldsymbol{\Theta}}) \leq \mathcal{K}_{\boldsymbol{X}}(\boldsymbol{A}, \widetilde{\boldsymbol{\Theta}}^s)$. Adapting the proof of Theorem 1 gives

$$\begin{aligned} \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 &\leq C'\rho_n\left(\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}^s) + \frac{\rho_n^2}{\gamma_n^2}\left(k^2 + n\log(k)\right)\right) \\ &\leq C'\rho_n\left(\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}^s) - \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \frac{\rho_n^2}{\gamma_n^2}\left(k^2 + n\log(k)\right)\right) \\ &\leq C'\rho_n\left(\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + 2\gamma_n n_s + \frac{\rho_n^2}{\gamma_n^2}\left(k^2 + n\log(k)\right)\right) \end{aligned} \tag{9}$$

where (9) follows from Lemma 20. Note that, if there exists a sequence γ_n such that $\rho_n = O(\gamma_n)$ and $n_s \leq \frac{k^2 \vee (n\log(k))}{\rho_n}$, the upper bound on the risk obtained in (9) matches the bound of Theorem 2 and is minimax optimal.

Without any assumption on the number of small entries of the matrix of connections probabilities, we choose $\gamma_n = \gamma(\rho_n) \triangleq n^{\frac{-2}{3}}\rho_n^{\frac{2}{3}}\left(k^2 + n\log(k)\right)^{\frac{1}{3}}$ and obtain the following bound.

Corollary 4. *Assume that $\boldsymbol{A}$ is drawn according to (1), and that $\rho_n = \omega(n^{-1})$. Let*

$$\begin{aligned} &\widehat{\boldsymbol{\Theta}}_{i<j} = \widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)}, \ \widehat{\boldsymbol{\Theta}}_{ii} = 0 \\ &(\widehat{\boldsymbol{Q}}, \widehat{z}) \in \underset{\boldsymbol{Q}\in[\gamma(\rho_n),\rho_n]^{k\times k}_{\mathrm{sym}}, z\in\mathcal{Z}_{n,k}}{\arg\min} \sum_{i<j} \boldsymbol{X}_{ij}\mathcal{K}(\boldsymbol{A}_{ij}, \boldsymbol{Q}_{z(i)z(j)}). \end{aligned}$$

There exists absolute constants $C, C' > 0$ such that with probability at least $1 - 9\exp\left(-C\rho_n\left(k^2 + n\log(k)\right)\right)$

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 \leq C'\rho_n\left(\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \rho_n^{\frac{2}{3}}n^{\frac{4}{3}}\left(k^2 + n\log(k)\right)^{\frac{1}{3}}\right).$$

If k is not too large, the rate of convergence is essentially multiplied by $(n\rho_n)^{\frac{2}{3}}$.

2.4 Choice of γ_n for sparse positive graphons

In Theorem 1 we have established an oracle bound for the maximum likelihood estimator with entries belonging to $[\gamma_n, \rho_n]$. Defining our estimator requires us to estimate the values of these two sparsity parameters, which are usually unknown. When matrix of connections probabilities $\boldsymbol{\Theta}^*$ is generated according to the sparse graphon model (4) where W^* is bounded away from 0, these bounds will be of the same order of magnitude and decrease as the expected node degree. Under this assumption, we can use $\widehat{d}$, the average number of edges, to estimate γ_n and ρ_n. Indeed, it is easy to see that, with probability close to 1, $\widehat{d}$ is close to $d = \rho_n\int_0^1\int_0^1 W^*(x,y)dxdy$, the expected node degree. Note that, if the graphon W^* is Hölder continuous or is a step function, assuming that $W^* > 0$ is enough to ensure that there exists a constant $C_{inf} > 0$ such that $W^* \geq C_{inf}$.

To simplify the exposition, we will assume that we observe all the entries of $\boldsymbol{A}$. Our results can be extended to the missing observations scheme described in Section 2.1 under the assumption that the entries of the sampling probability matrix $\boldsymbol{\Pi}$ are bounded away from 0. Let Ω be a subset of $\{(i,j) \in [n]^2, i < j\}$ of size n sampled independently of $\boldsymbol{A}$, and let

$$\begin{aligned} &\widehat{d} = \frac{1}{n}\sum_{(i,j)\in\Omega} \boldsymbol{A}_{ij} \\ &\widehat{\rho_n} = (\log(n))^{\frac{1}{5}}\widehat{d}\ , \ \widehat{\gamma_n} = (\log(n))^{-\frac{1}{5}}\widehat{d}. \end{aligned} \tag{10}$$

We use $\widehat{\rho_n}$ and $\widehat{\gamma_n}$ to build the restricted maximum likelihood estimator of the matrix of connections probabilities based on the the observations of $\boldsymbol{A}_{ij}$ with $\{(i,j) \in [n]^2, i<j\} \backslash \Omega$:

$$\begin{aligned} &\widehat{\boldsymbol{\Theta}}_{i<j} = \widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)}, \ \widehat{\boldsymbol{\Theta}}_{ii} = 0 \\ &(\widehat{\boldsymbol{Q}}, \widehat{z}) \in \underset{\boldsymbol{Q} \in [\widehat{\gamma_n}, \widehat{\rho_n}]^{k\times k}_{\mathrm{sym}}, z \in \mathcal{Z}_{n,k}}{\arg\min} \sum_{(i,j)\notin\Omega} \mathcal{K}(\boldsymbol{A}_{ij}, \boldsymbol{Q}_{z(i)z(j)}). \end{aligned} \tag{11}$$

We prove the following upper bound on the risk of this adaptive estimator:

Theorem 3. *Assume that $\boldsymbol{A}$ is drawn according to the sparse graphon model and $C_{inf} \triangleq \inf_{(x,y)\in[0,1]^2} W^*(x,y) > 0$, $\rho_n = o(\log(n)^{\frac{-1}{5}})$ and $\rho_n = \omega(n^{-1})$. Then, there exists positive constants N, C, C' depending only on C_{inf}, such that, for $n \geq N$, with probability at least $1 - 7\exp(-Cn\rho_n)$, we have*

$$\left\| \boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}} \right\|_2^2 \leq C' \rho_n \log(n) \left(\mathcal{K}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \left(k^2 + n\log(k)\right) \right).$$

In the sparse graphon model, if the graphon W^* is bounded away from 0 and $n^{-1} \ll \rho_n \ll \log(n)^{\frac{-1}{5}}$, our adaptive estimator is optimal in the minimax sense up to a log factor. When we can not assume that the graphon W^* is bounded away from 0, we can use the same trade-off as in (8) and choose $\widehat{\gamma_n} = \gamma(\widehat{\rho}_n)$. Then, with high probability, we obtain the following bound on the risk of the adaptative estimator:

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_2^2 \leq C' \rho_n \log(n) \left(\mathcal{K}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \left(\log(n)^{\frac{1}{5}} \rho_n\right)^{\frac{2}{3}} n^{\frac{4}{3}} \left(\frac{k^2}{n^2} + n\log(k) \right)^{\frac{1}{3}} \right).$$

2.5 Smooth graphons

We have established a non-asymptotic bound on the risk of the maximum likelihood estimator depending on the Kullback-Leibler divergence between $\boldsymbol{\Theta}^*$ and its oracle approximation by a block constant matrix corresponding to a SBM with k communities. While studying the graphon model (4), two classes of graphons are of particular interest: step function graphons and Hölder continuous graphons [45, 28, 20, 53]. A graphon W is called a *step function* if there exists a partition $S_1 \cup ... \cup S_k$ of $[0,1]$ into measurable sets such that the graphon W is constant on any product set $S_a \times S_b$. For step function graphons, the model corresponds to the stochastic block model described in (2): in this case, the oracle matrix $\widetilde{\boldsymbol{\Theta}}$ is equal to the matrix of connections probabilities $\boldsymbol{\Theta}^*$. Next, we bound the Kullback-Leibler divergence between $\boldsymbol{\Theta}^*$ and its oracle approximation by a block constant matrix for Hölder continuous graphons. We also provide the optimal choice for the number of communities k for our estimator.

We consider graphons that are weakly isomorphic to a smooth function. More precisely, for any $\alpha > 0$ and $M > 0$, let $\mathcal{F}_\alpha(M)$ be the class of Hölder functions, defined as follows:

$$\begin{aligned} \mathcal{F}_\alpha(M) = \Big\{ W \quad : \quad & [0,1]^2 \to [0,1], \ \forall (x,y), \ (x',y') \in [0,1]^2, \\ & \left| W(x',y') - \mathcal{P}_{\lfloor\alpha\rfloor}((x,y),(x'-x,y'-y)) \right| \leq M \left(|x-x'|^{\alpha - \lfloor\alpha\rfloor} + |y-y'|^{\alpha-\lfloor\alpha\rfloor} \right) \Big\} \end{aligned}$$

where $\mathcal{P}_{\lfloor\alpha\rfloor}((x,y),\cdot)$ is the Taylor polynomial of W of degree $\lfloor\alpha\rfloor$ at point (x,y). In particular, if $W \in \mathcal{F}_\alpha(M)$, $\forall (x,y), \ (x',y') \in [0,1]^2$,

$$\left| W(x',y') - W(x,y) \right| \leq M \left(|x-x'|^{\alpha\wedge 1} + |y-y'|^{\alpha\wedge 1} \right). \tag{12}$$

When the graphon is Hölder continuous, the following proposition provides an upper bound on the Kullback-Leibler divergence between $\boldsymbol{\Theta}^*$ and $\widetilde{\boldsymbol{\Theta}}$.

Proposition 1. *Consider the sparse graphon model (4) with $W^* \in \mathcal{F}_\alpha(M)$ where $\alpha, M > 0$ and we assume that $C_{inf} \triangleq \inf_{(x,y)\in[0,1]^2} W^*(x,y) > 0$, $\rho_n \leq 1 - C_{inf}$ and $\rho_n = \omega(n^{-1})$. Then, almost surely, there exists a k-block constant matrix $\boldsymbol{\Theta}^{bc}$ such that*

$$\mathcal{K}\left(\boldsymbol{\Theta}^*, \boldsymbol{\Theta}^{bc}\right) \leq \frac{4n^2 \rho_n M^2}{C_{inf}(1-\rho_n)} \left(\frac{1}{k}\right)^{2(\alpha\wedge 1)}. \tag{13}$$

Proposition 1 enables us to bound the bias of estimating $\boldsymbol{\Theta}^*$ by an oracle SBM with k communities. On the other hand, the bound given in Theorem 1 can be considered as the variance term of a block constant estimator with k blocks. To optimize the trade-off between these two terms, we choose k as follows

$$k = \left\lceil n^{\frac{1}{1+(\alpha\wedge 1)}} \rho_n^{\frac{1}{2+2(\alpha\wedge 1)}} \right\rceil \tag{14}$$

and obtain the following result:

Theorem 4. *Consider the sparse graphon model (4) with $W^* \in \mathcal{F}_\alpha(M)$ where $\alpha, M > 0$ and we assume that $C_{inf} \triangleq \inf_{(x,y)\in[0,1]^2} W^*(x,y) > 0$, $\rho_n \leq 1 - C_{inf}$ and that $\rho_n = \omega(n^{-1})$. Then, there exists constants $C, C' > 0$, depending only on M, α and C_{inf}, such that, the restricted maximum likelihood estimator defined by (6) constructed with k defined by (14) satisfies*

$$\left\|\boldsymbol{\Theta}^*_{ij} - \widehat{\boldsymbol{\Theta}}_{ij}\right\|_2^2 \leq C\rho_n \left(n^{\frac{2}{1+(\alpha\wedge 1)}} \rho_n^{\frac{1}{1+(\alpha\wedge 1)}} + n\log(\rho_n n)\right)$$

with probability larger than $1 - 9\exp(-C'\rho_n n \log(\rho_n n))$.

The bound obtained in Theorem 4 matches the minimax optimal rate established in [28] and proves that the maximum likelihood estimator is optimal for estimating the matrix of connections probabilities in graphon model for graphons W^* in the Hölder class.

3 Conclusion

We have studied the problem of estimating the matrix of connections probabilities for the inhomogeneous random graph model and the graphon model in the case of missing observations. We have established a non-asymptotic bound on the risk of the maximum likelihood estimator. In particular, we have shown that, if the entries of the probability matrix decrease at the same rate, our estimator achieves the minimax convergence rate. This result holds without requiring any knowledge on the probability of observing the entries of the adjacency matrix. When these probabilities are known, this convergence rate was already shown to be attained by the least square estimator, however this estimator cannot be computed in polynomial time and therefore it is not used in practise. While our estimator suffers from the same computational cost, its efficient approximations have been proposed in the literature, and have been implemented to study real life networks.

Acknowledgments

The work of O. Klopp was conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The authors want to thank Catherine Matias and Nicolas Verzelen for extremely valuable suggestions and discussions.

4 Proofs

The proof of Theorem 1 requires bounding the domain of definition of our estimator away from 0 and 1 in order to ensure that the loss function associated with the maximum likelihood estimator is Lipschitz. The Lipschitz constant here is equal to $\frac{1}{(1-\rho_n)\wedge\gamma_n}$. We balance this term by ρ_n by taking advantage of the sparsity of the graph, which implies, in particular, the low variance of $\boldsymbol{A}$. For ease of notations, we will assume that $1-\rho_n \leq \gamma_n$. This is the case when the graph is sparse, and our results still hold in the dense case if we replace γ_n by $\gamma_n \wedge (1-\rho_n)$ in our bounds.

4.1 Proof of Theorem 1

Let $\epsilon_n = C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k) + k^2\right)$ for some absolute constant C defined as the maximum of the absolute constants appearing in Lemma 6, Lemma 11 and Lemma 14, and let $\epsilon^0 \triangleq \rho_n\epsilon_n$. We start by considering the following two cases:

Case 1: $\|\widetilde{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 \leq 2\epsilon^0$. Then the statement of Theorem 1 follows from Lemma 19:

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 \leq 2\|\widetilde{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 + 2\|\boldsymbol{\Theta}^* - \widetilde{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 \leq 4\rho_n\epsilon_n + 16\rho_n\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}).$$

Case 2: $\|\widetilde{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 > 2\epsilon^0$. Then $\widehat{\boldsymbol{\Theta}}$ belongs to the set

$$\mathcal{S}_{\boldsymbol{\Pi}} = \left\{\boldsymbol{\Theta} \in \bigcup_{z\in\mathcal{Z}_{n,k}} \mathcal{T}_z : \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{2,\boldsymbol{\Pi}}^2 \geq 2\epsilon^0,\ \|\boldsymbol{\Theta}\|_\infty \leq \rho_n,\ \min_{i<j}\{\boldsymbol{\Theta}_{ij}\} \geq \gamma_n\right\}$$

and we use the following lemma.

Lemma 1. *There exists and absolute constant $C > 0$ such that for all $\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}$ simultaneously we have*

$$\left|\left\|\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{\Pi}}^2 - \left\|\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{X}}^2\right| \leq \frac{1}{2}\left\|\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{\Pi}}^2$$

with probability greater than $1 - 2\exp(-Cn\log(k))$.

Lemma 1 implies that with large probability, $\widehat{\boldsymbol{\Theta}}$ belongs to the set $\mathcal{S}_{\boldsymbol{X}}$ defined as

$$\mathcal{S}_{\boldsymbol{X}} = \left\{ \boldsymbol{\Theta} \in \bigcup_{z \in \mathcal{Z}_{n,k}} \mathcal{T}_z : \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{2,\boldsymbol{X}}^2 \geq \epsilon^0,\ \|\boldsymbol{\Theta}\|_\infty \leq \rho_n,\ \min_{i<j}\{\boldsymbol{\Theta}_{ij}\} \geq \gamma_n \right\}.$$

To bound $\|\widetilde{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2$ when $\widehat{\boldsymbol{\Theta}} \in \mathcal{S}_{\boldsymbol{X}} \cap \mathcal{S}_{\boldsymbol{\Pi}}$, we introduce the following notation. For $\boldsymbol{\Theta},\ \boldsymbol{\Theta}' \in (0,1)_{sym}^{n\times n}$ and $\boldsymbol{B}, \boldsymbol{C} \in [0,1]_{sym}^{n\times n}$ we set $\Delta\mathcal{K}_{\boldsymbol{B}}^{\boldsymbol{C}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}') = \mathcal{K}_{\boldsymbol{B}}(\boldsymbol{C}, \boldsymbol{\Theta}) - \mathcal{K}_{\boldsymbol{B}}(\boldsymbol{C}, \boldsymbol{\Theta}')$. Using Lemma 19 we get

$$\begin{aligned}
\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 &\leq 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widehat{\boldsymbol{\Theta}}) \\
&\leq 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + 8\rho_n \Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}).
\end{aligned}$$

On the other hand, the definition of $\widehat{\boldsymbol{\Theta}}$ implies that $\Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{A}}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}) \leq 0$ so

$$\begin{aligned}
\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 &\leq 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + 8\rho_n \Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}) - 8\rho_n \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{A}}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}) \qquad (15) \\
&\leq 8\rho_n \left(\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \left(\Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}) \right) + \left(\Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{A}}(\widehat{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Theta}}) \right) \right).
\end{aligned}$$

To bound the terms involved in equation (15), we control $\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}} \left| \Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) \right|$ using the concentration of $\boldsymbol{X}$ around its expectation $\boldsymbol{\Pi}$, and we control $\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}} \left| \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{A}}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) \right|$ conditionally on $\boldsymbol{X}$ using the concentration of $\boldsymbol{A}$ around its expectation $\boldsymbol{\Theta}^*$.

Lemma 2. *There exists absolute constants $C, C' > 0$ such that for all $\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}$ simultaneously we have*

$$\left| \Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) \right| \leq \frac{1}{2 \times 32\rho_n} \left\| \boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 + C\frac{\rho_n^2}{\gamma_n^2} \left(n\log(k) + k^2 \right)$$

with probability greater than $1 - 2\exp(-C'\rho_n n \log(k))$.

Lemma 3. *There exists absolute constants $C, C' > 0$ such that conditionally on $\boldsymbol{X}$, for all $\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{X}}$ simultaneously we have*

$$\left| \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{A}}(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) \right| \leq \frac{1}{4 \times 32\rho_n} \left\| \boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{X}}^2 + C\frac{\rho_n^2}{\gamma_n^2} \left(n\log(k) + k^2 \right)$$

with probability greater than $1 - 5\exp(-C'\rho_n n \log(k))$.

Combining Lemma 1, Lemma 2, Lemma 3 and (15) yields that there exists two absolute constants $C, C' > 0$ such that with probability greater than $1 - 9\exp\left(-C'\rho_n n \log(k)\right)$

$$\begin{aligned}
\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 &\leq 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + 8\rho_n \times \frac{1}{2 \times 32\rho_n} \left\| \widehat{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 + 8\rho_n \times \frac{1}{4 \times 32\rho_n} \left\| \widehat{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{X}}^2 \\
&\quad + C\rho_n \frac{\rho_n^2}{\gamma_n^2} \left(n\log(k) + k^2 \right) \\
&\leq 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \frac{1}{8} \left\| \widehat{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 + \frac{1}{16} \times \frac{3}{2} \left\| \widehat{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 + C\rho_n \frac{\rho_n^2}{\gamma_n^2} \left(n\log(k) + k^2 \right) \\
&\leq 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \frac{1}{2} \left\| \boldsymbol{\Theta}^* - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 + \frac{1}{2} \left\| \boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 + C\rho_n \frac{\rho_n^2}{\gamma_n^2} \left(n\log(k) + k^2 \right). \qquad (16)
\end{aligned}$$

Lemma 19 and (16) imply that there exists two absolute constants $C, C' > 0$ such that with probability larger than $1 - 9\exp\left(-C'\rho_n n \log(k)\right)$,

$$\frac{1}{2}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_{2,\boldsymbol{\Pi}}^2 \leq 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + \frac{1}{2} \times 8\rho_n \mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + C\rho_n \frac{\rho_n^2}{\gamma_n^2} \left(n\log(k) + k^2 \right).$$

This completes the proof of Theorem 1.

4.2 Proof of Lemma 1

To prove Lemma 1, we show that the probability of the following "bad" event is small:

$$\mathcal{E} \triangleq \left\{ \exists \boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}} : \left| \left\| \boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 - \left\| \boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{X}}^2 \right| > \frac{1}{2} \left\| \boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}} \right\|_{2,\boldsymbol{\Pi}}^2 \right\}.$$

We use a standard peeling argument (see, e.g., [27]): we slice $\mathcal{S}_{\mathbf{\Pi}}$ in different sets, on which we control $\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\mathbf{\Pi}}^2$. Recall that $\epsilon_n \triangleq C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right)$ where the absolute constant C is larger than the constant appearing in Lemma 6, and that $\epsilon^0 \triangleq \rho_n\epsilon_n$. For $l\in\mathbb{N}^*$, we set

$$\mathcal{S}_{l,\mathbf{\Pi}} \triangleq \left\{\mathbf{\Theta}\in\mathcal{S}_{\mathbf{\Pi}} : 2^{l-1}(2\epsilon^0) \le \left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\mathbf{\Pi}}^2 \le 2^l(2\epsilon^0)\right\}.$$

If the event $\mathcal{E}$ holds, there exists $l\in\mathbb{N}^*$ such that $\mathbf{\Theta}\in\mathcal{S}_{l,\mathbf{\Pi}}$ and

$$\left|\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\mathbf{\Pi}}^2 - \left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\boldsymbol{X}}^2\right| > \frac{1}{2}\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\mathbf{\Pi}}^2.$$

Note that $\mathbb{E}\left[\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\boldsymbol{X}}^2\right] = \left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\mathbf{\Pi}}^2$. The events that we need to control are the following:

$$\mathcal{E}_l \triangleq \left\{\exists\mathbf{\Theta}\in\mathcal{S}_{l,\mathbf{\Pi}} : \left|\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\boldsymbol{X}}^2 - \mathbb{E}\left[\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\boldsymbol{X}}^2\right]\right| > \frac{2^{l-1}(2\epsilon^0)}{2}\right\}.$$

If $\mathcal{E}$ holds for some $\mathbf{\Theta}\in\mathcal{S}_{\mathbf{\Pi}}$, there exists $l\in\mathbb{N}^*$ such that $\mathbf{\Theta}\in\mathcal{S}_{l,\mathbf{\Pi}}$, thus there exists $l\in\mathbb{N}^*$ such that $\mathcal{E}_l$ holds, i.e., $\mathcal{E}\subset\bigcup_{l\in\mathbb{N}^*}\mathcal{E}_{l,\mathbf{\Pi}}$. For $T>0$, let $\mathcal{S}_{\mathbf{\Pi}}(T)$ be defined as follows:

$$\mathcal{S}_{\mathbf{\Pi}}(T) = \left\{\mathbf{\Theta}\in\bigcup_{z\in\mathcal{Z}_{n,k}}\mathcal{T}_z :\ \|\mathbf{\Theta}\|_\infty\le\rho_n,\ \min_{i<j}\{\mathbf{\Theta}_{ij}\}\ge\gamma_n, \left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\mathbf{\Pi}}^2\le T\right\}.$$

We see that $\mathcal{S}_{l,\mathbf{\Pi}}\subset\mathcal{S}_{\mathbf{\Pi}}(2^l\epsilon^0)$, so we only need to control the probability of the events

$$\mathcal{E}(T) = \left\{\exists\mathbf{\Theta}\in\mathcal{S}_{\mathbf{\Pi}}(T) : \left|\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\boldsymbol{X}}^2 - \mathbb{E}\left[\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\boldsymbol{X}}^2\right]\right| > \frac{T}{4}\right\}.$$

The following lemma helps us bound the probability of the events $\mathcal{E}(T)$.

Lemma 4. *For $T>\epsilon^0$, let $Z_T = \sup_{\mathbf{\Theta}\in\mathcal{S}_{\mathbf{\Pi}}(T)}\left|\left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\mathbf{\Pi}}^2 - \left\|\mathbf{\Theta}-\widetilde{\mathbf{\Theta}}\right\|_{2,\boldsymbol{X}}^2\right|$. There exists an absolute constant $C>0$ such that*

$$\mathbb{P}\left(Z_T\ge\frac{T}{4}\right)\le\exp\left(-\frac{CT}{\rho_n}\right).$$

Proof. To prove Lemma 4, we first show that Z_T concentrates around its expectation and then bound this term.

Lemma 5. *Let Z_T be defined as in 4. Then*

$$\mathbb{P}\left(Z_T > 2\mathbb{E}[Z_T]+\frac{T}{16}\right)\le\exp\left(-\frac{T}{64\rho_n}\right).$$

Lemma 6. *Let Z_T be as in Lemma 4, then there exists an absolute constant $C>0$ such that*

$$\mathbb{E}\left[Z_T\right]\le\frac{T}{16}+C\rho_n\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right).$$

Putting together Lemma 5 and Lemma 6, we get that there exists an absolute constant $C>0$ such that

$$\mathbb{P}\left(Z_T\ge\frac{3T}{16}+\frac{C}{8}\rho_n\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right)\right)\le\exp\left(-\frac{T}{64\rho_n}\right).$$

Our choice of ϵ_0 allows us to conclude that for $T\ge2\epsilon_0$, $\frac{C}{8}\rho_n\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right)\le\frac{T}{16}$ and

$$\mathbb{P}\left(Z_T\ge\frac{T}{4}\right)\le\exp(-\frac{T}{64\rho_n}).$$

□

For this choice of ϵ_0,

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &\leq \sum_{l=1}^{\infty} \mathbb{P}\left(\mathcal{E}(2^l(2\epsilon^0))\right) \\ &\leq \sum_{l=1}^{\infty} \exp\left(-2C\epsilon^0 2^l/\rho_n\right) \\ &\leq \sum_{l=1}^{\infty} \exp\left(-2Cl\log(2)\epsilon^0/\rho_n\right) \\ &\leq \frac{\exp\left(-2C\log(2)\epsilon^0/\rho_n\right)}{1-\exp\left(-2C\log(2)\epsilon^0/\rho_n\right)} = \frac{1}{\exp\left(-2C\log(2)\epsilon^0/\rho_n\right)-1} \leq 2\exp\left(-Cn\log(k)\right)\end{aligned}$$

for n large enough. This completes the proof of Lemma 1.

4.2.1 Proof of Lemma 5

To control the deviation of Z_T from its expectation, we apply the following theorem from Bousquet, as stated in [21], Theorem 3.3.16.

Theorem 5 (Bousquet)**.** *Let X_i, $i \in \mathbb{N}$ be independent $\mathcal{S}$-valued random variables, and let $\mathcal{F}$ be a countable class of functions $f = (f_1, ..., f_n) : \mathcal{S} \to [-1,1]^n$ such that $\mathbb{E}[f_i(X_i)] = 0$ for all $f \in \mathcal{F}$ and $i \in [n]$. Set $Z = \sup_{f\in\mathcal{F}} \left| \sum_{1\leq i\leq n} f_i(X_i) \right|$ and $v = \sup_{f\in\mathcal{F}} \sum_{1\leq i\leq n} \mathbb{E}\left[f_i(X_i)^2\right]$. Then, for all $x > 0$,*

$$\mathbb{P}\left(Z > \mathbb{E}[Z] + \frac{x}{3} + \sqrt{2x(2\mathbb{E}[Z]+v)}\right) \leq \exp(-x).$$

We apply Theorem 5 to the random variable

$$\begin{aligned} Z_T &= \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)} \left| \left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{\Pi}}^2 - \left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{X}}^2 \right| \\ &= \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)} \left| \sum_{1\leq i<j\leq n} (\boldsymbol{\Pi}_{ij} - \boldsymbol{X}_{ij}) \left(\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right)^2 \right| \\ &= \rho_n \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)} \left| \sum_{1\leq i<j\leq n} f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij}) \right| \end{aligned}$$

where we set $f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij}) \triangleq \frac{(\boldsymbol{X}_{ij}-\boldsymbol{\Pi}_{ij})(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij})^2}{\rho_n}$. The set of functions $\left\{f_{ij}^{\boldsymbol{\Theta}}, \boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}(T)\right\}$ is separable and we can apply Theorem 5 (see, e.g., [21], Section 2.1). Note that for all $1 \leq i < j \leq n$, $\mathbb{E}\left[f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij})\right] = 0$, $\left|f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij})\right| \leq 1$, $\mathbb{E}\left[(\boldsymbol{X}_{ij}-\boldsymbol{\Pi}_{ij})^2\right] \leq \boldsymbol{\Pi}_{ij}$ and $\left|\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right| \leq \rho_n$ so

$$\begin{aligned} \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)} \sum_{1\leq i<j\leq n} \mathbb{E}\left[f_{ij}^{\boldsymbol{\Theta}}(X_{ij})^2\right] &\leq \frac{1}{\rho_n^2} \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)} \sum_{1\leq i<j\leq n} \boldsymbol{\Pi}_{ij} \left(\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right)^4 \\ &\leq \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)} \sum_{1\leq i<j\leq n} \boldsymbol{\Pi}_{ij} \left(\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right)^2 \\ &\leq T. \end{aligned}$$

Theorem 5 implies that

$$\begin{aligned} \mathbb{P}\left(\frac{Z_T}{\rho_n} > \frac{1}{\rho_n}\mathbb{E}[Z_T] + \frac{x}{3} + \sqrt{2x\left(\frac{2}{\rho_n}\mathbb{E}[Z_T]+T\right)}\right) &\leq \exp(-x) \\ \mathbb{P}\left(Z_T > \mathbb{E}[Z_T] + \frac{\rho_n x}{3} + \sqrt{2x\rho_n(2\mathbb{E}[Z_T]+\rho_n T)}\right) &\leq \exp(-x) \\ \mathbb{P}\left(Z_T > \mathbb{E}[Z_T] + \frac{\rho_n x}{3} + 2\rho_n x + \mathbb{E}[Z_T] + \rho_n\sqrt{2xT}\right) &\leq \exp(-x) \end{aligned}$$

where we have used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $2\sqrt{ab} \leq a + b$. Setting $x = \frac{T}{64\rho_n}$ and noticing that $\rho_n \leq 1$ leads to

$$\mathbb{P}\left(Z_T > 2\mathbb{E}[Z_T] + \frac{T}{16}\right) \leq \exp(-\frac{T}{64\rho_n}).$$

4.2.2 Proof of Lemma 6

Once we have bounded Z_T by its expectation, we bound $\mathbb{E}[Z_T]$. To do so, we use a symetrization argument and Talagrand's contraction principle (see, e.g., [21] for a proof).

Lemma 7 (Symmetrization). *Let $\{\boldsymbol{Y}_i\}_{1\leq i\leq n}$ be independent random variables, $\{\boldsymbol{\epsilon}_i\}_{1\leq i\leq n}$ be a Rademacher sequence, and $\mathcal{A}$ be a subset of $\mathbb{R}^n$, then*

$$\mathbb{E}\left[\sup_{\boldsymbol{A}\in\mathcal{A}}\left|\sum_{1\leq i\leq n}(\boldsymbol{Y}_i-\mathbb{E}[\boldsymbol{Y}_i])\boldsymbol{A}_i\right|\right]\leq 2\mathbb{E}\left[\sup_{\boldsymbol{A}\in\mathcal{A}}\left|\sum_{1\leq i\leq n}\boldsymbol{\epsilon}_i\boldsymbol{Y}_i\boldsymbol{A}_i\right|\right].$$

Lemma 8 (Talagrand's contraction principle). *Let $\{\phi_i\}_{1\leq i\leq n}:\mathbb{R}\to\mathbb{R}$ be 1-Lipshitz functions vanishing at 0, $\mathcal{A}$ be a compact subset of $\mathbb{R}^n$ and $\{\boldsymbol{\epsilon}_i\}_{1\leq i\leq n}$ be a Rademacher sequence, then*

$$\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{A}}\left|\sum_{1\leq i\leq n}\boldsymbol{\epsilon}_i\phi_i\left(\boldsymbol{\Theta}_i\right)\right|\right]\leq 2\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{A}}\left|\sum_{1\leq i\leq n}\boldsymbol{\epsilon}_i\boldsymbol{\Theta}_i\right|\right].$$

Recall that

$$\mathbb{E}[Z_T] \;=\; \mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\leq i<j\leq n}(\boldsymbol{\Pi}_{ij}-\boldsymbol{X}_{ij})\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)^2\right|\right].$$

Let $(\boldsymbol{\epsilon}_{ij})_{1\leq i<j\leq n}$ be a Rademacher sequence. Lemma 7 implies

$$\mathbb{E}[Z_T] \;\leq\; 2\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\leq i<j\leq n}\boldsymbol{\epsilon}_{ij}\boldsymbol{X}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)^2\right|\right].$$

For any $1\leq i<j\leq n$, let $\phi_{ij}:x\to\frac{x^2}{2\rho_n}$. Note that on $[-\rho_n,\rho_n]$, ϕ_{ij} is a 1-Lipschitz and vanishes at 0. Applying Lemma 8, we get that

$$\begin{aligned}\mathbb{E}[Z_T] &\leq 4\rho_n\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\leq i<j\leq n}\boldsymbol{\epsilon}_{ij}\phi_{ij}\left(\boldsymbol{X}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)\right)\right|\right]\\ &\leq 8\rho_n\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\leq i<j\leq n}\boldsymbol{\epsilon}_{ij}\boldsymbol{X}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)\right|\right]. \qquad (17)\end{aligned}$$

We bound the term in using the following lemma.

Lemma 9. *Let $\boldsymbol{B}\in\{\boldsymbol{\Pi},\boldsymbol{X}\}$ and let $\boldsymbol{\Sigma}$ be a random matrix such that almost surely, $\|\boldsymbol{\Sigma}\|_\infty\leq 1$ and that conditionally on $\boldsymbol{B}$, for all $1\leq i<j\leq n$ the coefficients $\boldsymbol{\Sigma}_{ij}$ are independent and centered. Assume that there exists $\alpha>0$ such that for all $\boldsymbol{\Theta}\in\mathbb{R}^{n\times n}_{sym}$, $\sum_{i<j}\mathbb{E}^{\boldsymbol{B}}\left[\boldsymbol{\Sigma}_{ij}^2\boldsymbol{\Theta}_{ij}^2\right]\leq\alpha\|\boldsymbol{\Theta}\|_{2,\boldsymbol{B}}^2$. There exists an absolute constant C such that*

$$\mathbb{E}^{\boldsymbol{B}}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{B}}(T)}\left|\sum_{1\leq i<j\leq n}\boldsymbol{\Sigma}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)\right|\right]\leq\frac{\gamma_n\alpha T}{32\times 64^2\rho_n^2}+C\frac{\rho_n^2}{\gamma_n}(n\log(k)+k^2).$$

Note that for all $1\leq i<j\leq n$, $\mathbb{E}[\boldsymbol{X}_{ij}^2]\leq\boldsymbol{\Pi}_{ij}$, so for all $\boldsymbol{\Theta}\in\mathbb{R}^{n\times n}_{sym}$, $\sum_{i<j}\mathbb{E}\left[\boldsymbol{\epsilon}_{ij}^2\boldsymbol{X}_{ij}^2\boldsymbol{\Theta}_{ij}^2\right]\leq\|\boldsymbol{\Theta}\|_{2,\boldsymbol{\Pi}}^2$. We apply Lemma 9 with $\boldsymbol{B}=\boldsymbol{\Pi}$, $\alpha=1$ and for all $1\leq i<j\leq n$, $\boldsymbol{\Sigma}_{ij}=\boldsymbol{\epsilon}_{ij}\boldsymbol{X}_{ij}$ and combine it with (17) to get that for some absolute constant C

$$\begin{aligned}\mathbb{E}[Z_T] &\leq \frac{T\gamma_n\times 8\rho_n}{32\times 64^2\rho_n^2}+C\rho_n\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right)\\ &\leq \frac{T}{4\times 64}+C\rho_n\frac{\rho_n^2}{\gamma_n}\left(n\log(k)+k^2\right).\end{aligned}$$

This concludes the proof of Lemma 6.

4.2.3 Proof of Lemma 9

To get an upper bound on $\mathbb{E}^{\boldsymbol{B}}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{B}}(T)}\left|\sum_{1\le i<j\le n}\boldsymbol{\Sigma}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)\right|\right]$, we use Bernstein's inequality, which we state here for the reader's convenience:

Theorem 6 (Bernstein's inequality)**.** *Let $X_1,...,X_n$ be independent centered random variables. Assume that for all $i\in[n]$, $|X_i|\le M$ almost surely, then*

$$\mathbb{P}\left(\left|\sum_{1\le i\le n}X_i\right|\ge\sqrt{2t\sum_{1\le i\le n}\mathbb{E}[X_i^2]}+\frac{2M}{3}t\right)\le 2e^{-t}.$$

Recall that for $\boldsymbol{B}\in\{\boldsymbol{\Pi},\boldsymbol{X}\}$,

$$\mathcal{S}_{\boldsymbol{B}}(T)=\left\{\boldsymbol{\Theta}\in\bigcup_{z\in\mathcal{Z}_{n,k}}\mathcal{T}_z:\ \|\boldsymbol{\Theta}\|_\infty\le\rho_n,\ \min_{i<j}\{\boldsymbol{\Theta}_{ij}\}\ge\gamma_n,\left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{B}}^2\le T\right\}$$

and let $\mathcal{S}_z(T)\triangleq\mathcal{T}_z\cap\mathcal{S}_{\boldsymbol{B}}(T)$ be the set of matrices in $\mathcal{S}_{\boldsymbol{B}}(T)$ that are block constant for the label z. Let $\widetilde{\boldsymbol{\Theta}}^z$ be the projection of $\widetilde{\boldsymbol{\Theta}}$ onto $\mathcal{T}_z$ for the $\boldsymbol{B}$-weighted Frobenius norm:

$$\widetilde{\boldsymbol{\Theta}}^z\triangleq\operatorname*{arg\,min}_{\boldsymbol{\Theta}\in\mathcal{T}_z}\left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{B}}.$$

Note that if $\mathcal{S}_z(T)\neq\emptyset$, then $\widetilde{\boldsymbol{\Theta}}^z\in\mathcal{S}_z(T)$. If $\mathcal{S}_z(T)=\emptyset$, we set $\sup_{\boldsymbol{\Theta}\in\mathcal{S}_z(T)}\left|\left\langle\boldsymbol{\Sigma}\middle|\widetilde{\boldsymbol{\Theta}}^z-\boldsymbol{\Theta}\right\rangle\right|=0$. We decompose the error in two terms.

$$\begin{aligned}\mathbb{E}^{\boldsymbol{B}}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{B}}(T)}\left|\sum_{1\le i<j\le n}\boldsymbol{\Sigma}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)\right|\right]&\le\mathbb{E}^{\boldsymbol{B}}\left[\sup_{z\in\mathcal{Z}_{n,k},\ \mathcal{S}_z(T)\neq\emptyset}\left|\left\langle\boldsymbol{\Sigma}\middle|\widetilde{\boldsymbol{\Theta}}-\widetilde{\boldsymbol{\Theta}}^z\right\rangle\right|\right]\\&\quad+\ \mathbb{E}^{\boldsymbol{B}}\left[\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{\Theta}\in\mathcal{S}_z(T)}\left|\left\langle\boldsymbol{\Sigma}\middle|\widetilde{\boldsymbol{\Theta}}^z-\boldsymbol{\Theta}\right\rangle\right|\right]\\&\le(I)+(II).\end{aligned}\tag{18}$$

The term (I) denotes $\mathbb{E}^{\boldsymbol{B}}\left[\sup_{z\in\mathcal{Z}_{n,k},\ \mathcal{S}_z(T)\neq\emptyset}\left|\left\langle\boldsymbol{\Sigma}\middle|\widetilde{\boldsymbol{\Theta}}-\widetilde{\boldsymbol{\Theta}}^z\right\rangle\right|\right]$ and corresponds to the error induced by an error on the label. The term (II) denotes $\mathbb{E}^{\boldsymbol{B}}\left[\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{\Theta}\in\mathcal{S}_z(T)}\left|\left\langle\boldsymbol{\Sigma}\middle|\widetilde{\boldsymbol{\Theta}}^z-\boldsymbol{\Theta}\right\rangle\right|\right]$ and corresponds to the error induced by a Bernoulli noise.

Control of (I): To control the first term of (18), recall that for any $z\in\mathcal{Z}_{n,k}$ such that $\mathcal{S}_z(T)\neq\emptyset$, $\widetilde{\boldsymbol{\Theta}}^z\in\mathcal{S}_z(T)$ and by hypothesis, $\sum_{i<j}\mathbb{E}^{\boldsymbol{B}}\left[\boldsymbol{\Sigma}_{ij}^2\left(\widetilde{\boldsymbol{\Theta}}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}^z\right)^2\right]\le\alpha\left\|\widetilde{\boldsymbol{\Theta}}-\widetilde{\boldsymbol{\Theta}}^z\right\|_{2,\boldsymbol{B}}^2\le\alpha T$. Furthermore, $\|\boldsymbol{\Sigma}\|_\infty\le 1$ so $\left|\boldsymbol{\Sigma}_{ij}(\widetilde{\boldsymbol{\Theta}}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}^z)\right|\le\rho_n$. Since $|\mathcal{Z}_{n,k}|\le n\log(k)$, the union bound and Bernstein's inequality imply

$$\begin{aligned}\mathbb{P}^{\boldsymbol{B}}\left(\sup_{z\in\mathcal{Z}_{n,k},\ \mathcal{S}_z(T)\neq\emptyset}\left|\left\langle\boldsymbol{\Sigma}|\widetilde{\boldsymbol{\Theta}}-\widetilde{\boldsymbol{\Theta}}^z\right\rangle\right|\ge\sqrt{2\alpha T\left(t+n\log(k)\right)}+\frac{2\rho_n}{3}\left(t+n\log(k)\right)\right)&\le 2e^{-t}\\\mathbb{P}^{\boldsymbol{B}}\left(\sup_{z\in\mathcal{Z}_{n,k},\ \mathcal{S}_z(T)\neq\emptyset}\left|\left\langle\boldsymbol{\Sigma}|\widetilde{\boldsymbol{\Theta}}-\widetilde{\boldsymbol{\Theta}}^z\right\rangle\right|\ge\frac{\gamma_n\alpha T}{64^3\rho_n^2}+\left(\frac{2\rho_n}{3}+\frac{64^3\rho_n^2}{\gamma_n}\right)\left(t+n\log(k)\right)\right)&\le 2e^{-t}.\end{aligned}$$

Integrating the last inequality and using $\frac{\rho_n}{\gamma_n}\ge 1$, we get that for some absolute constant C

$$(I)\le\frac{\alpha\gamma_n T}{64^3\rho_n^2}+C\frac{\rho_n^2}{\gamma_n}n\log(k).\tag{19}$$

Control of (II): The control of the second term of (18) is more involved. We adapt the argument developped in [28] and consider only $z\in\mathcal{Z}_{n,k}$ such that $\mathcal{S}_z(T)\neq\emptyset$. By property of the projection, we have for all $\boldsymbol{\Theta}\in\mathcal{S}_z(T)$, $\left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}^z\right\|_{2,\boldsymbol{B}}^2\le\left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{B}}^2\le T$. Thus $\left(\widetilde{\boldsymbol{\Theta}}^z-\boldsymbol{\Theta}\right)\in\mathcal{A}_z(T)$, where

$$\mathcal{A}_z(T)\triangleq\left\{\boldsymbol{\Theta}\in\mathcal{T}_z:\|\boldsymbol{\Theta}\|_\infty\le\rho_n,\ \|\boldsymbol{\Theta}\|_{2,\boldsymbol{B}}^2\le T\right\},$$

so $\sup_{\boldsymbol{\Theta}\in\mathcal{S}_z(T)}\left|\left\langle\boldsymbol{\Sigma}|\widetilde{\boldsymbol{\Theta}}^z-\boldsymbol{\Theta}\right\rangle\right|\leq\sup_{\boldsymbol{\Theta}\in\mathcal{A}_z(T)}|\langle\boldsymbol{\Sigma}|\boldsymbol{\Theta}\rangle|$. Let $\widehat{\boldsymbol{T}}^z\in\mathcal{A}_z(T)$ be such that

$$\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{T}}^z\right\rangle\right|\triangleq\sup_{\boldsymbol{\Theta}\in\mathcal{A}_z(T)}|\langle\boldsymbol{\Sigma}|\boldsymbol{\Theta}\rangle|\,. \tag{20}$$

Note that $\boldsymbol{\Theta}\to|\langle\boldsymbol{\Sigma}|\boldsymbol{\Theta}\rangle|$ is continous and reaches its supremum on $\mathcal{A}_z(T)$. Indeed, either for all $1\leq i<j\leq n$, $\boldsymbol{B}_{ij}>0$ so $\|\cdot\|_{\boldsymbol{B}}$ is a norm and $\mathcal{A}_z(T)$ is compact, or we can find a subspace $\mathcal{V}$ of $\mathbb{R}^{\frac{(n-1)(n-2)}{2}}$ of dimension $|\{1\leq i<j\leq n:\boldsymbol{B}_{ij}>0\}|$ such that for all $\boldsymbol{\Theta}\in\mathbb{R}^{\frac{(n-1)(n-2)}{2}}$, $\langle\boldsymbol{\Sigma}|\boldsymbol{\Theta}\rangle=\langle\boldsymbol{\Sigma}|\mathcal{P}_{\mathcal{V}}(\boldsymbol{\Theta})\rangle$ where $\mathcal{P}_{\mathcal{V}}$ denotes the projection onto $\mathcal{V}$, and $\mathcal{A}_z(T)\cap\mathcal{V}$ is compact.

To control $\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{T}}^z\right\rangle\right|$, we build a finite set with small cardinality that approximates $\widehat{\boldsymbol{T}}^z$ well both in the weighted Frobenius norm and in the supremum norm. More precisely, our goal is to construct a finite set $\widetilde{\mathcal{C}}_z(T)$ containing a matrix $\widehat{\boldsymbol{V}}$ such that $2\left(\widehat{\boldsymbol{T}}^z-\widehat{\boldsymbol{V}}\right)\in\mathcal{A}_z(T)$. To apply Bernstein's inequality, we also need to be able to control the suppremum norm on this set. Our first step will be to construct such a set.

We denote by $\mathcal{B}_r$ the ball centered at $\mathbf{0}$ and of radius r for the weighted Frobenius norm $\|\cdot\|_{2,\boldsymbol{B}}$. Let $\mathcal{C}_z$ be a minimal $\sqrt{T}/2$-net for the weighted Frobenius norm on $\mathcal{B}_{\sqrt{T}}\cap\mathcal{T}_z$. Note that $\mathcal{A}_z\subset\mathcal{B}_{\sqrt{T}}\cap\mathcal{T}_z$, so there exists $\widehat{\boldsymbol{V}}\in\mathcal{C}_z(T)$ such that $\left\|\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{T}}^z\right\|_{2,\boldsymbol{B}}\leq\frac{\sqrt{T}}{2}$. Since our choice of net does not allow us to directly bound $\left\|\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{T}}^z\right\|_{\infty}$, we extend this net using the following argument. For any $\boldsymbol{V}\in\mathcal{C}_z$ and any matrix $\boldsymbol{U}\in\{-1,0,1\}^{k\times k}$, let $\boldsymbol{V}^{\boldsymbol{U}}\in\mathbb{R}^{n\times n}$ be such that $\boldsymbol{V}^{\boldsymbol{U}}_{ii}=0$ and for all $i<j$,

$$\boldsymbol{V}^{\boldsymbol{U}}_{ij}=\operatorname{sign}(\boldsymbol{V}_{ij})\left(|\boldsymbol{V}_{ij}|\wedge\rho_n\right)\left(1-\left|\boldsymbol{U}_{z(i)z(j)}\right|\right)+\boldsymbol{U}_{z(i)z(j)}\frac{\rho_n}{2}.$$

Recall that $\left\|\widehat{\boldsymbol{T}}^z\right\|_{\infty}\leq\rho_n$ so for any $\boldsymbol{V}\in\mathcal{C}_z(T)$ we have $\left|\operatorname{sign}(\boldsymbol{V}_{ij})\left(|\boldsymbol{V}_{ij}|\wedge\rho_n\right)-\widehat{\boldsymbol{T}}^z_{ij}\right|\leq\left|\boldsymbol{V}_{ij}-\widehat{\boldsymbol{T}}^z_{ij}\right|$. This implies that $\left\|\boldsymbol{V}^{\mathbf{0}}-\widehat{\boldsymbol{T}}^z\right\|_{2,\boldsymbol{B}}\leq\left\|\boldsymbol{V}-\widehat{\boldsymbol{T}}^z\right\|_{2,\boldsymbol{B}}$.

Now, let $\widetilde{\mathcal{C}}_z(T)=\left\{\boldsymbol{V}^{\boldsymbol{U}}:\boldsymbol{V}\in\mathcal{C}_z(T),\boldsymbol{U}\in\{-1,0,1\}^{k\times k}_{sym}\right\}$ and $\widehat{\boldsymbol{U}}=\operatorname*{arg\,min}_{\boldsymbol{U}\in\{-1,0,1\}^{k\times k}}\left\|\widehat{\boldsymbol{V}}^{\boldsymbol{U}}-\widehat{\boldsymbol{T}}^z\right\|_{\infty}$. By definition, for all $(a,b)\in k\times k$, $\widehat{\boldsymbol{U}}$ minimises $\left|\widehat{\boldsymbol{V}}^{\boldsymbol{U}}_{z^{-1}(a)z^{-1}(b)}-\widehat{\boldsymbol{T}}^z_{z^{-1}(a)z^{-1}(b)}\right|$, so it is also a minimizer of $\left\|\widehat{\boldsymbol{V}}^{\boldsymbol{U}}-\widehat{\boldsymbol{T}}^z\right\|_{2,\boldsymbol{B}}=\sum_{a,b\in[k]}\left(\sum_{(i,j)\in z^{-1}(a)\times z^{-1}(b),i\neq j}\boldsymbol{B}_{ij}\right)\left|\widehat{\boldsymbol{V}}^{\boldsymbol{U}}_{z^{-1}(a)z^{-1}(b)}-\widehat{\boldsymbol{T}}^z_{z^{-1}(a)z^{-1}(b)}\right|^2$. Therefore

$$\left\|\widehat{\boldsymbol{V}}^{\widehat{U}}-\widehat{\boldsymbol{T}}^z\right\|_{2,\boldsymbol{B}}\leq\left\|\widehat{\boldsymbol{V}}^{\mathbf{0}}-\widehat{\boldsymbol{T}}^z\right\|_{2,\boldsymbol{B}}\leq\left\|\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{T}}^z\right\|_{2,\boldsymbol{B}}\leq\frac{\sqrt{T}}{2}.$$

Furthermore $\left\|\widehat{\boldsymbol{V}}^{\widehat{U}}-\widehat{\boldsymbol{T}}^z\right\|_{\infty}\leq\left\|\widehat{\boldsymbol{V}}^{U^*}-\widehat{\boldsymbol{T}}^z\right\|_{\infty}$, where $\boldsymbol{U}^*_{ab}=\operatorname{sign}(\widehat{\boldsymbol{T}}^z_{z^{-1}(a)z^{-1}(b)})$. By construction,

$$\left\|\widehat{\boldsymbol{V}}^{U^*}-\widehat{\boldsymbol{T}}^z\right\|_{\infty}=\sup_{i<j}\left|\widehat{\boldsymbol{T}}^z_{ij}-\operatorname{sign}(\widehat{\boldsymbol{T}}^z_{ij})\frac{\rho_n}{2}\right|=\sup_{i<j}\left|\left|\widehat{\boldsymbol{T}}^z_{ij}\right|-\frac{\rho_n}{2}\right|\leq\frac{\rho_n}{2}.$$

Hence, $2\left(\widehat{\boldsymbol{T}}^z-\widehat{\boldsymbol{V}}^{\hat{U}}\right)\in\mathcal{A}_z(T)$. Thus, we have shown that

$$\begin{aligned}
2\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{T}}^z-\widehat{\boldsymbol{V}}^{\hat{U}}\right\rangle\right| &\leq \sup_{\boldsymbol{\Theta}\in\mathcal{A}_z(T)}|\langle\boldsymbol{\Sigma}|\boldsymbol{\Theta}\rangle|\triangleq\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{T}}^z\right\rangle\right|\\
2\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{T}}^z\right\rangle\right|-2\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{V}}^{\hat{U}}\right\rangle\right| &\leq \left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{T}}^z\right\rangle\right|\\
\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{T}}^z\right\rangle\right| &\leq 2\left|\left\langle\boldsymbol{\Sigma}|\widehat{\boldsymbol{V}}^{\hat{U}}\right\rangle\right|.
\end{aligned}$$

This and (20) allows us to conclude that

$$\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{\Theta}\in\mathcal{S}_z(T)}\left|\left\langle\boldsymbol{\Sigma}|\widetilde{\boldsymbol{\Theta}}^z-\boldsymbol{\Theta}\right\rangle\right|\leq 2\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{V}\in\widetilde{\mathcal{C}}_z(T)}|\langle\boldsymbol{\Sigma}|\boldsymbol{V}\rangle|\,. \tag{21}$$

To bound the right hand side of (21), we recall that by hypothesis for any $\boldsymbol{V} \in \widetilde{\mathcal{C}}_z(T)$, $\sum_{i<j} \mathbb{E}^{\boldsymbol{B}}\left[\boldsymbol{\Sigma}_{ij}^2 \boldsymbol{V}_{ij}^2\right] \leq \alpha \|\boldsymbol{V}\|_{2,\boldsymbol{B}}^2$ and note that $\|\boldsymbol{V}\|_\infty \leq \rho_n$ and $\|\boldsymbol{V}\|_{2,\boldsymbol{B}} \leq \sqrt{T}$. We use Bernstein's inequality and the union bound to obtain

$$\mathbb{P}\left(\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{V}\in\widetilde{\mathcal{C}}_z(T)} |\langle\boldsymbol{\Sigma}|\boldsymbol{V}\rangle| \geq \sqrt{2\alpha T t} + \frac{2}{3}t\right) \quad \leq \quad 2e^{-t+n\log(k)+\sup_{\boldsymbol{V}\in\widetilde{\mathcal{C}}_z(T)}\log\left(\left|\widetilde{\mathcal{C}}_z(T)\right|\right)}. \tag{22}$$

By construction of $\widetilde{\mathcal{C}}_z(T)$, we have $\left|\widetilde{\mathcal{C}}_z(T)\right| = |\mathcal{C}_z(T)| \times 3^{k^2}$. The following classical result on the covering number of a ball will help us bound $|\mathcal{C}_z(T)|$ (see, e.g., Lemma 5.2 in [49]).

Lemma 10. *Let $\mathcal{B}_r$ the ball of a subspace of $\mathbb{R}^n$ of dimension d centered at $\boldsymbol{0}$ and of radius r for the euclidean norm, and $\mathcal{N}(\mathcal{B}_r, \epsilon)$ its ϵ-covering number, that is the minimal cardinality of a set $\mathcal{C}$ such that for all $\boldsymbol{X} \in \mathcal{B}_r$, there exists $\boldsymbol{Y} \in \mathcal{C}$ such that $\|\boldsymbol{X} - \boldsymbol{Y}\| \leq \epsilon$. Then*

$$\mathcal{N}(\mathcal{B}_r, \epsilon) \leq \left(\frac{3r}{\epsilon}\right)^d.$$

Extending the proof Lemma 10 to a weighed euclidean norm is straightforward. Putting Lemma 10 into equation (22) and noting that $\mathcal{T}_z$ spans a subspace of $\mathbb{R}^{\frac{(n-1)(n-2)}{2}}$ of dimension $\frac{k(k-1)}{2}$, we get that for some absolute constant C

$$\mathbb{P}\left(\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{V}\in\widetilde{\mathcal{C}}_z(T)} |\langle\boldsymbol{\Sigma}|\boldsymbol{V}\rangle| \geq \sqrt{2\alpha T\left(t + n\log(k) + k^2\log(C)\right)} + \frac{2\rho_n}{3}\left(t + n\log(k) + k^2\log(C)\right)\right) \quad \leq \quad 2e^{-t}$$

$$\mathbb{P}\left(\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{V}\in\widetilde{\mathcal{C}}_z(T)} |\langle\boldsymbol{\Sigma}|\boldsymbol{V}\rangle| \geq \frac{\alpha\gamma_n T}{2\times 64^3\rho_n^2} + \left(\frac{2\rho_n}{3} + \frac{2\times 64^2\rho_n^2}{\gamma_n}\right)\left(t + n\log(k) + k^2\log(C)\right)\right) \leq 2e^{-t}.$$

We integrate and find for some absolute constant $C > 0$

$$\mathbb{E}\left[\sup_{z\in\mathcal{Z}_{n,k}}\sup_{\boldsymbol{\Theta}\in\mathcal{S}_z(T)}\left|\left\langle\boldsymbol{\Sigma}|\widetilde{\boldsymbol{\Theta}}^z - \boldsymbol{\Theta}\right\rangle\right|\right] \leq \frac{\alpha\gamma_n T}{64^3\rho_n^2} + C\frac{\rho_n^2}{\gamma_n}\left(n\log(k) + k^2\right). \tag{23}$$

Combining the bounds (23) and (19) yields the desired result.

4.3 Proof of Lemma 2

The proof of Lemma 2 closely follows that of Lemma 1 and we only sketch it. Recall that $\epsilon_n \triangleq C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k) + k^2\right)$ where the absolute constant C is larger than the constant appearing in Lemma 11, and that $\epsilon^0 \triangleq \rho_n\epsilon_n$. We show that the probability of the following "bad" event is small:

$$\mathcal{E} \triangleq \left\{\exists\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}} : \left|\Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right| > \frac{1}{2\times 32\rho_n}\left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{\Pi}}^2 + \epsilon_n\right\}.$$

Again, we slice $\mathcal{S}_{\boldsymbol{\Pi}}$ in different sets $\mathcal{S}_{l,\boldsymbol{\Pi}}$ defined as $\mathcal{S}_{l,\boldsymbol{\Pi}} \triangleq \left\{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}} : 32^{l-1}(2\epsilon^0) \leq \left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{\Pi}}^2 \leq 32^l(2\epsilon^0)\right\}$ on which we control the events $\mathcal{E}_l \triangleq \left\{\exists\boldsymbol{\Theta}\in\mathcal{S}_{l,\boldsymbol{\Pi}} : \left|\Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right| > \frac{32^{l-1}\times 2\epsilon^0}{4\times 32\rho_n} + \epsilon_n\right\}$. To do this, we set $\mathcal{S}_{\boldsymbol{\Pi}}(T) \triangleq \left\{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}} : \left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|_{2,\boldsymbol{\Pi}}^2 \leq T\right\}$ and we control the probability of the events

$$\mathcal{E}(T) = \left\{\exists\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T) : \left|\Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right| > \frac{T}{64^2\rho_n} + \epsilon_n\right\}.$$

The following lemma helps us bound the probability of the events $\mathcal{E}(T)$.

Lemma 11. *Let $\widetilde{Z}_T = \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\Delta\mathcal{K}_{\boldsymbol{\Pi}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right|$. There exists two absolute constants C, $C' > 0$ such that*

$$\mathbb{P}\left(\widetilde{Z}_T \geq \frac{T}{64^2\rho_n} + C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k) + k^2\right)\right) \leq \exp\left(-\frac{C'T\gamma_n^2}{\rho_n^2}\right).$$

Proof. To prove Lemma 11, we first show that Z_T concentrates around its expectation and then bound this term.

Lemma 12. *Let $\widetilde{Z}_T$ be defined as in Lemma 11. Then there exists an absolute constant $C > 0$ such that*

$$\mathbb{P}\left(\widetilde{Z}_T > 2\mathbb{E}[\widetilde{Z}_T] + \frac{T}{2 \times 64^2 \rho_n}\right) \leq \exp\left(-\frac{CT\gamma_n^2}{\rho_n^2}\right).$$

Lemma 13. *Let $\widetilde{Z}_T$ be as in Lemma 11, then there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\left[\widetilde{Z}_T\right] \leq \frac{T}{4 \times 64^2 \rho_n} + C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k) + k^2\right). \tag{24}$$

Putting together Lemma 12 and Lemma 13, we get that there exists two absolute constants $C, C' > 0$ such that

$$\mathbb{P}\left(\widetilde{Z}_T \geq \frac{T}{64^2\rho_n} + C\frac{\rho_n^2}{\gamma_n^2}(n\log(k) + k^2)\right) \leq \exp\left(-\frac{C'T\gamma_n^2}{\rho_n^2}\right).$$

This concludes the proof of Lemma 11. □

Lemma 11 and the arguments developped to prove Lemma 1 help us conclude the proof of Lemma 2.

4.3.1 Proof of Lemma 12

Recall that by definition of $\widetilde{Z}_T$,

$$\begin{aligned}
\widetilde{Z}_T &= \sup_{\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}(T)} \left| \sum_{1 \leq i < j \leq n} (\boldsymbol{\Pi}_{ij} - \boldsymbol{X}_{ij}) \left(\boldsymbol{\Theta}^* \log\left(\frac{\widetilde{\boldsymbol{\Theta}}}{\boldsymbol{\Theta}}\right) + (1 - \boldsymbol{\Theta}^*) \log\left(\frac{1 - \widetilde{\boldsymbol{\Theta}}}{1 - \boldsymbol{\Theta}}\right)\right)\right| \\
&= \frac{1}{\gamma_n} \sup_{\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}(T)} \left| \sum_{1 \leq i < j \leq n} f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij}) \right|
\end{aligned}$$

where we set $f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij}) \triangleq \gamma_n (\boldsymbol{\Pi}_{ij} - \boldsymbol{X}_{ij}) \left(\boldsymbol{\Theta}^*_{ij} \log\left(\frac{\widetilde{\boldsymbol{\Theta}}_{ij}}{\boldsymbol{\Theta}_{ij}}\right) + (1 - \boldsymbol{\Theta}^*_{ij}) \log\left(\frac{1-\widetilde{\boldsymbol{\Theta}}_{ij}}{1-\boldsymbol{\Theta}_{ij}}\right)\right)$. Assuming that $\gamma_n \leq 1 - \rho_n$, $x \to \log(x)$ and $x \to \log(1-x)$ are $\frac{1}{\gamma_n}$ - Lipshitz on $[\gamma_n, \rho_n]$ so

$$\left| \boldsymbol{\Theta}^*_{ij} \log\left(\frac{\widetilde{\boldsymbol{\Theta}}_{ij}}{\boldsymbol{\Theta}_{ij}}\right) + (1 - \boldsymbol{\Theta}^*_{ij}) \log\left(\frac{1 - \widetilde{\boldsymbol{\Theta}}_{ij}}{1 - \boldsymbol{\Theta}_{ij}}\right)\right| \leq \boldsymbol{\Theta}^* \frac{\left|\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right|}{\gamma_n} + (1 - \boldsymbol{\Theta}^*) \frac{\left|\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right|}{\gamma_n} \leq \frac{\left|\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right|}{\gamma_n}$$

which implies that for all $1 \leq i < j \leq n$, $\left|f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij})\right| \leq 1$. Moreover for all $1 \leq i < j \leq n$, $\mathbb{E}\left[f_{ij}^{\boldsymbol{\Theta}}(\boldsymbol{X}_{ij})\right] = 0$ and $\mathbb{E}\left[(\boldsymbol{X}_{ij} - \boldsymbol{\Pi}_{ij})^2\right] \leq \boldsymbol{\Pi}_{ij}$, hence

$$\begin{aligned}
\sup_{\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}(T)} \sum_{1 \leq i < j \leq n} \mathbb{E}\left[f_{ij}^{\boldsymbol{\Theta}}(X_{ij})^2\right] &\leq \gamma_n^2 \sup_{\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}(T)} \sum_{1 \leq i \leq n} \boldsymbol{\Pi}_{ij} \left(\boldsymbol{\Theta}^*_{ij} \log\left(\frac{\widetilde{\boldsymbol{\Theta}}_{ij}}{\boldsymbol{\Theta}_{ij}}\right) + (1 - \boldsymbol{\Theta}^*_{ij}) \log\left(\frac{1 - \widetilde{\boldsymbol{\Theta}}_{ij}}{1 - \boldsymbol{\Theta}_{ij}}\right)\right)^2 \\
&\leq \gamma_n^2 \sup_{\boldsymbol{\Theta} \in \mathcal{S}_{\boldsymbol{\Pi}}(T)} \sum_{1 \leq i \leq n} \boldsymbol{\Pi}_{ij} \frac{1}{\gamma_n^2} \left(\boldsymbol{\Theta}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right)^2 \\
&\leq T.
\end{aligned}$$

Then, Theorem 5 implies

$$\begin{aligned}
\mathbb{P}\left(\gamma_n \widetilde{Z}_T > \gamma_n \mathbb{E}[\widetilde{Z}_T] + \frac{x}{3} + \sqrt{2x(2\gamma_n \mathbb{E}[\widetilde{Z}_T] + T)}\right) &\leq \exp(-x) \\
\mathbb{P}\left(\widetilde{Z}_T > \mathbb{E}[\widetilde{Z}_T] + \frac{x}{3\gamma_n} + \frac{4x}{\gamma_n} + \mathbb{E}[\widetilde{Z}_T] + \frac{2x \times 4 \times 64^2 \rho_n}{\gamma_n^2} + \frac{T}{4 \times 64^2 \rho_n}\right) &\leq \exp(-x) \\
\mathbb{P}\left(\widetilde{Z}_T > 2\mathbb{E}[\widetilde{Z}_T] + \frac{9 \times 64^2 x \rho_n}{\gamma_n^2} + \frac{T}{4 \times 64^2 \rho_n}\right) &\leq \exp(-x)
\end{aligned}$$

where we have used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $\sqrt{ab} \leq a + b$ and $\frac{\rho_n}{\gamma_n} \geq 1$. Setting $x = \frac{T\gamma_n^2}{9 \times 64^2 \times 4 \times 64^2 \rho_n^2}$ yields the desired result.

4.3.2 Proof of Lemma 13

In Lemma 13, we bound

$$\mathbb{E}\left[\widetilde{Z}_T\right] = \mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\le i<j\le n}(\boldsymbol{\Pi}_{ij}-\boldsymbol{X}_{ij})\left(\boldsymbol{\Theta}^*_{ij}\log\left(\frac{\widetilde{\boldsymbol{\Theta}}_{ij}}{\boldsymbol{\Theta}_{ij}}\right)+(1-\boldsymbol{\Theta}^*)\log\left(\frac{1-\widetilde{\boldsymbol{\Theta}}_{ij}}{1-\boldsymbol{\Theta}_{ij}}\right)\right)\right|\right].$$

Let $(\boldsymbol{\epsilon}_{ij})_{1\le i<j\le n}$ be a Rademacher sequence. We apply Lemma 7 and get

$$\mathbb{E}\left[\widetilde{Z}_T\right] \le 2\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\le i<j\le n}\boldsymbol{\epsilon}_{ij}\boldsymbol{X}_{ij}\left(\boldsymbol{\Theta}^*_{ij}\log\left(\frac{\widetilde{\boldsymbol{\Theta}}_{ij}}{\boldsymbol{\Theta}_{ij}}\right)+(1-\boldsymbol{\Theta}^*_{ij})\log\left(\frac{1-\widetilde{\boldsymbol{\Theta}}_{ij}}{1-\boldsymbol{\Theta}_{ij}}\right)\right)\right|\right].$$

For any $1\le i<j\le n$, let $\phi_{ij}: x\to\frac{\gamma_n}{2\rho_n}\boldsymbol{X}_{ij}\left(\boldsymbol{\Theta}^*_{ij}\log\left(\frac{\widetilde{\boldsymbol{\Theta}}_{ij}-x}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right)+(1-\boldsymbol{\Theta}^*_{ij})\log\left(\frac{1+x-\widetilde{\boldsymbol{\Theta}}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)$. Note that on $[\widetilde{\boldsymbol{\Theta}}_{ij}-\rho_n,\widetilde{\boldsymbol{\Theta}}_{ij}-\gamma_n]$, ϕ_{ij} is 1-Lipschitz and vanishes at 0. Then we apply Lemma 8 and compute

$$\begin{aligned}\mathbb{E}\left[\widetilde{Z}_T\right] &\le \frac{4\rho_n}{\gamma_n}\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\le i<j\le n}\boldsymbol{\epsilon}_{ij}\phi_{ij}\left(\boldsymbol{X}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)\right)\right|\right]\\ &\le \frac{8\rho_n}{\gamma_n}\mathbb{E}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{\Pi}}(T)}\left|\sum_{1\le i<j\le n}\boldsymbol{\epsilon}_{ij}\boldsymbol{X}_{ij}\left(\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right)\right|\right].\end{aligned}$$

Now, applying Lemma 9 with $\alpha=1$ and $\boldsymbol{B}=\boldsymbol{\Pi}$ allows us to conclude that there exists an absolute constant $C>0$ such that

$$\mathbb{E}\left[\widetilde{Z}_T\right]\le\frac{T}{8\times64^2\rho_n}+C\frac{\rho_n^3}{\gamma_n^2}(n\log(n)+k^2)\le\frac{T}{8\times64^2\rho_n}+C\frac{\rho_n^2}{\gamma_n^2}(n\log(n)+k^2).$$

4.4 Proof of Lemma 3

The proof of Lemma 3 closely follows that of Lemma 2, and we only sketch it. Recall that $\epsilon_n\triangleq C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right)$ where the absolute constant C is larger than the constant appearing in Lemma 14, and that $\epsilon^0\triangleq\rho_n\epsilon_n$. We show that conditionally on $\boldsymbol{X}$, the probability of the following "bad" event is small and does not depend on $\boldsymbol{X}$:

$$\mathcal{E}_{\boldsymbol{X}}\triangleq\left\{\exists\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}:\left|\Delta\mathcal{K}^{\boldsymbol{\Theta}^*}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})-\Delta\mathcal{K}^{\boldsymbol{A}}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right|>\frac{1}{2\times64\rho_n}\left\|\boldsymbol{\Theta}-\boldsymbol{\Theta}\right\|^2_{2,\boldsymbol{X}}+\epsilon_n\right\}.$$

We slice $\mathcal{S}_{\boldsymbol{X}}$ in the following sets $\mathcal{S}_{l,\boldsymbol{X}}\triangleq\left\{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}:64^{l-1}\epsilon^0\le\left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|^2_{2,\boldsymbol{X}}\le64^l\epsilon^0\right\}$ and control the probability of the events $\mathcal{E}_{l,\boldsymbol{X}}\triangleq\left\{\exists\boldsymbol{\Theta}\in\mathcal{S}_{l,\boldsymbol{X}}:\left|\Delta\mathcal{K}^{\boldsymbol{\Theta}^*}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})-\Delta\mathcal{K}^{\boldsymbol{A}}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right|>\frac{64^l\epsilon^0}{2\times64^2\rho_n}+\epsilon_n\right\}$. To do this, we control the probability of the events $\mathcal{E}_{\boldsymbol{X}}(T)=\left\{\exists\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T):\left|\Delta\mathcal{K}^{\boldsymbol{\Theta}^*}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})-\Delta\mathcal{K}^{\boldsymbol{A}}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right|>\frac{T}{2\times64^2\rho_n}+\epsilon_n\right\}$ where $\mathcal{S}_{\boldsymbol{X}}(T)=\left\{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}:\left\|\boldsymbol{\Theta}-\widetilde{\boldsymbol{\Theta}}\right\|^2_{2,\boldsymbol{X}}\le T\right\}$.

Lemma 14. *Let $Z_{T,\boldsymbol{X}}=\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)}\left|\Delta\mathcal{K}^{\boldsymbol{\Theta}^*}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})-\Delta\mathcal{K}^{\boldsymbol{A}}_{\boldsymbol{X}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right|$. There exists two absolute constants C, $C'>0$ such that $\mathbb{P}^{\boldsymbol{X}}\left(Z_{T,\boldsymbol{X}}\ge\frac{T}{2\times64^2\rho_n}+C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right)\right)\le4\exp\left(-\frac{C'\gamma_n^2T}{\rho_n^2}\right)$.*

Proof. To prove Lemma 14, we first show that $Z_{T,\boldsymbol{X}}$ concentrates around its expectation and then bound this term.

Lemma 15. *Let $Z_{T,\boldsymbol{X}}$ be as in Lemma 14, then there exists two absolute constants C, $C'>0$ such that*

$$\mathbb{P}^{\boldsymbol{X}}\left(\left|Z_{T,\boldsymbol{X}}-\mathbb{E}^{\boldsymbol{X}}(Z_{T,\boldsymbol{X}})\right|>\frac{C\rho_n}{\gamma_n^2}+\frac{T}{4\times64^2\rho_n}\right)\le4\exp\left(-\frac{C'T\gamma_n^2}{\rho_n^2}\right).$$

Lemma 16. *Let $Z_{T,\boldsymbol{X}}$ be as in Lemma 14, then there exists an absolute constant $C>0$ such that*

$$\mathbb{E}^{\boldsymbol{X}}\left[Z_{T,\boldsymbol{X}}\right]\le\frac{T}{4\times64^2\rho_n}+\frac{C\rho_n^2}{\gamma_n^2}\left(n\log(k)+k^2\right).$$

Putting together Lemma 15 and Lemma 16, we get that

$$\mathbb{P}^{\boldsymbol{X}}\left(Z_{T,\boldsymbol{X}} \geq \frac{T}{2\times 64^2\rho_n} + C\frac{\rho_n^2}{\gamma_n^2}\left(n\log(k) + \frac{1}{\rho_n} + k^2\right)\right) \leq 4\exp\left(-\frac{C'\gamma_n^2 T}{\rho_n^2}\right).$$

If $n\rho_n \to \infty$, for n large enough $n > \frac{1}{\rho_n}$. This yields the desired result. □

We combine Lemma 14 and the arguments developed in Lemma 1, and note that $\mathbb{P}^{\boldsymbol{X}}(\mathcal{E}_{\boldsymbol{X}})$ does not depend on $\boldsymbol{X}$ to conclude the proof of Lemma 3.

4.4.1 Proof of Lemma 15

In this Section, we prove the Lemma 15 that helps us bound $\left|Z_{T,\boldsymbol{X}} - \mathbb{E}^{\boldsymbol{X}}(Z_{T,\boldsymbol{X}})\right|$ with hight probability. To prove that $Z_{T,\boldsymbol{X}}$ concentrates around its mean, we use the following version of Talagrand's Theorem for Lipschitz convex functions (for a proof, see Theorem 3.3 of [14]).

Theorem 7. *Suppose that $f : [-1,1]^N \to \mathbb{R}$ is a convex Lipschitz function with Lipschitz constant L. Let $R_1, ..., R_N$ be independent random variables taking value in $[-1,1]$. Let $Z := f(R_1, ..., R_N)$. Then for any $t \geq 0$,*

$$\mathbb{P}\left(|Z - \mathbb{E}(Z)| > 16L + t\right) \leq 4e^{\left(\frac{-t^2}{2L^2}\right)}.$$

Recall that

$$\begin{aligned} Z_{T,\boldsymbol{X}} &= \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left|\Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{\Theta}^*}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}}) - \Delta\mathcal{K}_{\boldsymbol{X}}^{\boldsymbol{A}}(\boldsymbol{\Theta},\widetilde{\boldsymbol{\Theta}})\right| \\ &= \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left|\sum_{1\leq i<j\leq n} (\boldsymbol{X}_{ij}\boldsymbol{A}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\Theta}^*_{ij})\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|. \end{aligned}$$

Note that $Z_{T,\boldsymbol{X}} = f(\boldsymbol{A})$ where $f(\boldsymbol{R})$ is defined for $\boldsymbol{R} \in [-1,1]^{\frac{(n-1)(n-2)}{2}}$ by

$$f : \boldsymbol{R} \to \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left|\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}(\boldsymbol{R}_{ij} - \boldsymbol{\Theta}^*_{ij})\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|.$$

It is easy to see that f is indeed convex. Our next step is to show that f is Lipschitz. Let $\boldsymbol{R}, \boldsymbol{S} \in [-1,1]^{\frac{(n-1)(n-2)}{2}}$,

$$\begin{aligned} |f(\boldsymbol{R}) - f(\boldsymbol{S})| &= \left|\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left|\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}(\boldsymbol{R}_{ij} - \boldsymbol{\Theta}^*_{ij})\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|\right. \\ &\quad \left. - \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left|\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}(\boldsymbol{S}_{ij} - \boldsymbol{\Theta}^*_{ij})\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|\right| \\ &\leq \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left|\left|\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}(\boldsymbol{R}_{ij} - \boldsymbol{\Theta}^*_{ij})\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|\right. \\ &\quad \left. - \left|\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}(\boldsymbol{S}_{ij} - \boldsymbol{\Theta}^*_{ij})\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|\right| \\ &\leq \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left|\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}(\boldsymbol{R}_{ij} - \boldsymbol{S}_{ij})\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right| \\ &\leq \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left\{\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}\left|(\boldsymbol{R}_{ij} - \boldsymbol{S}_{ij})\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right| + \boldsymbol{X}_{ij}\left|(\boldsymbol{R}_{ij} - \boldsymbol{S}_{ij})\log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right|\right\} \\ &\leq \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \|\boldsymbol{R} - \boldsymbol{S}\|_2 \left(\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}\left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)^2 + \boldsymbol{X}_{ij}\left(\log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)^2\right)^{\frac{1}{2}} \end{aligned}$$

where we have used that $\boldsymbol{X} \in \{0,1\}^{n\times n}$. Thus f is Lipschitz with Lipschitz constant

$$\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left(\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij} \left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)^2 + \boldsymbol{X}_{ij} \left(\log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)^2 \right)^{\frac{1}{2}}.$$

As stated before, assuming that $\gamma_n \leq 1-\rho_n$, $x \to \log(x)$ and $x \to \log(1-x)$ are Lipschitz functions on $[\gamma_n, \rho_n]$ with Lipschitz constant γ_n^{-1}. Thus f is Lipschitz with Lipschitz constant

$$\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left(\sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij} \left(\frac{\left|\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right|}{\gamma_n}\right)^2 + \boldsymbol{X}_{ij} \left(\frac{\left|\boldsymbol{\Theta}_{ij}-\widetilde{\boldsymbol{\Theta}}_{ij}\right|}{\gamma_n}\right)^2 \right)^{\frac{1}{2}}.$$

This implies

$$|f(\boldsymbol{R})-f(\boldsymbol{S})| \quad \leq \quad \|\boldsymbol{R}-\boldsymbol{S}\|_2 \sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \frac{\sqrt{2}\left\|\widetilde{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\right\|_{2,\boldsymbol{X}}}{\gamma_n} \leq \|\boldsymbol{R}-\boldsymbol{S}\|_2 \frac{\sqrt{2T}}{\gamma_n}.$$

We have shown that f has a Lipschitz constant $L = \frac{\sqrt{2T}}{\gamma_n}$. Applying Theorem 7 for $t = \frac{T}{8\times 64^2\rho_n}$, we get

$$\mathbb{P}^{\boldsymbol{X}}\left(\left| Z_{T,\boldsymbol{X}} - \mathbb{E}(Z_{T,\boldsymbol{X}})\right| > \frac{16\sqrt{2T}}{\gamma_n} + \frac{T}{8\times 64^2\rho_n}\right) \leq 4\exp\left(\frac{-T\gamma_n^2}{8^2\times 2\times 64^4\rho_n^2}\right).$$

Using for $\beta>0$, $2\sqrt{ab} \leq \beta a^2 + b^2/\beta$ yields

$$\mathbb{P}^{\boldsymbol{X}}\left(\left| Z_{T,\boldsymbol{X}} - \mathbb{E}(Z_{T,\boldsymbol{X}})\right| > \frac{8^2\times 64^2\times 16\rho_n}{\gamma_n^2} + \frac{T}{8\times 64^2\rho_n} + \frac{T}{8\times 64^2\rho_n}\right) \leq 4\exp\left(\frac{-T\gamma_n^2}{4\times 4^2\times 32^4\rho_n^2}\right).$$

This concludes the proof of Lemma 15.

4.5 Proof of Lemma 16

Once we have shown that $Z_{T,\boldsymbol{X}}$ concentrates around its mean, we bound $\mathbb{E}[Z_{T,\boldsymbol{X}}]$. To do so, we follow the steps of Lemma 13. Let $\boldsymbol{\epsilon}_{1\leq i<j\leq n}$ a Rademacher sequence. Applying Lemma 7, we get

$$\begin{aligned}\mathbb{E}^{\boldsymbol{X}}[Z_{T,\boldsymbol{X}}] &= \mathbb{E}^{\boldsymbol{X}}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left| \sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}(\boldsymbol{A}_{ij}-\mathbb{E}[\boldsymbol{A}_{ij}]) \left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|\right] \\ &\leq 2\mathbb{E}^{\boldsymbol{X}}\left[\sup_{\boldsymbol{\Theta}\in\mathcal{S}_{\boldsymbol{X}}(T)} \left| \sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}\boldsymbol{\epsilon}_{ij}\boldsymbol{A}_{ij} \left(\log\left(\frac{\boldsymbol{\Theta}_{ij}}{\widetilde{\boldsymbol{\Theta}}_{ij}}\right) - \log\left(\frac{1-\boldsymbol{\Theta}_{ij}}{1-\widetilde{\boldsymbol{\Theta}}_{ij}}\right)\right)\right|\right].\end{aligned}$$

For any $1\leq i<j\leq n$, let $\phi_{ij} : x \to \frac{1}{2}\gamma_n \boldsymbol{X}_{ij}\boldsymbol{A}_{ij}\left(\log\left(\frac{\widetilde{\Theta}_{ij}-x}{\widetilde{\Theta}_{ij}}\right) - \log\left(\frac{1+x-\widetilde{\Theta}_{ij}}{1-\widetilde{\Theta}_{ij}}\right)\right)$. Note that ϕ_{ij} is 1-Lipschitz and vanishes at 0 on the interval $[\widetilde{\boldsymbol{\Theta}}_{ij}-\rho_n, \widetilde{\boldsymbol{\Theta}}_{ij}-\gamma_n]$. Indeed,

$$\begin{aligned}\phi_i(x)' &= \frac{1}{2}\gamma_n \boldsymbol{X}_{ij}\boldsymbol{A}_{ij}\left(\frac{-1}{\widetilde{\boldsymbol{\Theta}}_{ij}-x} - \frac{1}{1+x-\widetilde{\boldsymbol{\Theta}}_{ij}}\right) \\ &\leq \boldsymbol{X}_{ij}\boldsymbol{A}_{ij}.\end{aligned}$$

By definition of the functions ϕ_{ij},

$$\mathbb{E}^{\boldsymbol{X}}[Z_{T,\boldsymbol{X}}] \leq \frac{4}{\gamma_n}\mathbb{E}^{\boldsymbol{X}}\left[\sup_{\boldsymbol{\Theta}\in S_{\boldsymbol{X}}(T)} \left| \sum_{i<j} \boldsymbol{\epsilon}_{ij}\phi_{ij}(\boldsymbol{X}_{ij}\boldsymbol{A}_{ij}(\widetilde{\boldsymbol{\Theta}}_{ij}-\boldsymbol{\Theta}_{ij}))\right|\right].$$

We apply Lemma 8 to get

$$\mathbb{E}^{\boldsymbol{X}}[Z_{T,\boldsymbol{X}}] \leq \frac{8}{\gamma_n}\mathbb{E}^{\boldsymbol{X}}\left[\sup_{\boldsymbol{\Theta}\in S_{\boldsymbol{X}}(T)} \left| \sum_{i<j} \boldsymbol{X}_{ij}\boldsymbol{\epsilon}_{ij}\boldsymbol{A}_{ij}(\widetilde{\boldsymbol{\Theta}}_{ij}-\boldsymbol{\Theta}_{ij}))\right|\right]. \tag{25}$$

Next, we apply Lemma 9 with $\boldsymbol{B} = \boldsymbol{X}$, $\boldsymbol{\Sigma}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{A}_{ij}\boldsymbol{\epsilon}_{ij}$ and $\alpha = \rho_n$. Note that $\|\boldsymbol{\Sigma}\|_\infty \leq 1$ and that for any matrix $\boldsymbol{\Theta}$, $\sum_{i<j}\mathbb{E}^{\boldsymbol{X}}\left[\boldsymbol{X}_{ij}^2\boldsymbol{\epsilon}_{ij}^2\boldsymbol{\Theta}_{ij}^2\right] \leq \rho_n \|\boldsymbol{\Theta}\|_{2,\boldsymbol{X}}^2$. Combining Lemma 9 and (25) yields

$$\mathbb{E}^{\boldsymbol{X}}\left[Z_{T,\boldsymbol{X}}\right] \leq \frac{8}{\gamma_n} \times \left(\frac{T\gamma_n}{32 \times 64^2 \rho_n} + C\frac{\rho_n^2}{\gamma_n}\left(n\log(k) + k^2\right)\right).$$

This concludes the proof of Lemma 16.

4.6 Proof of Theorem 3

Our proof relies on two steps: first, we show that with high probability, $\widehat{d}$ is close to its expected value, which belongs to $[\gamma_n, \rho_n]$. More precisely, let $\underline{\gamma_n} = \frac{C_{inf}}{2}\rho_n \log(n)^{\frac{-1}{5}}$ and $\underline{\rho_n} = \left(1 + \frac{C_{inf}}{2}\right)\rho_n \log(n)^{\frac{1}{5}}$. We prove that with high probability, $\underline{\gamma_n} \leq \widehat{\gamma_n} \leq \gamma_n$ and $\rho_n \leq \widehat{\rho_n} \leq \underline{\rho_n}$. Then this implies that the oracle matrix $\widetilde{\boldsymbol{\Theta}}$ belongs to the set of definition of our estimator and its likelihood is greater than that of $\widehat{\boldsymbol{\Theta}}$. Then both $\widetilde{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}$ belong to the set $\left[\underline{\gamma_n}, \underline{\rho_n}\right]^{n\times n}$ and we adapt the proof of Theorem 1 to get the desired result.

Lemma 17. *Let $\mathcal{E} = \left\{\widehat{\gamma_n} \in \left[\underline{\gamma_n}, \gamma_n\right], \widehat{\rho_n} \in \left[\rho_n, \underline{\rho_n}\right]\right\}$. There exists a positive constant C and an integer N, both depending only on C_{inf}, such that $\forall n \geq N$, $\mathbb{P}(\mathcal{E}) \geq 1 - \exp(-Cn\rho_n)$.*

Proof. Note that $\|\boldsymbol{A} - \boldsymbol{\Theta}^*\|_\infty \leq 1$ almost surely, and that for all $1 \leq i < j \leq n$, $(\boldsymbol{A}_{ij} - \boldsymbol{\Theta}^*_{ij})$ is centered and has a variance smaller than ρ_n. Applying Bernstein's inequality 6 yields

$$\mathbb{P}\left(\left|\sum_{(i,j)\in\Omega}\left(\boldsymbol{A}_{ij} - \boldsymbol{\Theta}^*_{ij}\right)\right| \geq \sqrt{2tn\rho_n} + \frac{3t}{2}\right) \quad \leq \quad 2e^{-t},\ \forall t > 0.$$

Choosing $t = \rho_n nC$ with $C > 0$ sucht that $\sqrt{2C} + \frac{3C}{2} \leq \frac{C_{inf}}{2}$ yields

$$\mathbb{P}\left(\left|\widehat{d} - \frac{\sum_{(i,j)\in\Omega}\boldsymbol{\Theta}^*_{ij}}{n}\right| \geq \frac{C_{inf}}{2}\rho_n\right) \quad \leq \quad 2e^{-Cn\rho_n}.$$

Note that in the sparse graphon model (4), when $0 < C_{inf} \triangleq \inf_{(x,y)\in[0,1]^2} W^*(x,y)$, we see that $\gamma_n = C_{inf}\rho_n$ and $\frac{\sum_{(i,j)\in\Omega}\boldsymbol{\Theta}^*_{ij}}{n} \in [\gamma_n, \rho_n] = [C_{inf}\rho_n, \rho_n]$. So, with probability greater than $1 - 2e^{-Cn\rho_n}$, $\widehat{A} \in \left[\frac{C_{inf}\rho_n}{2}, (1 + \frac{C_{inf}}{2})\rho_n\right]$. Let N be such that $\log(N)^{-\frac{1}{5}} \leq \frac{C_{inf}}{1+\frac{C_{inf}}{2}}$ and $\log(N)^{\frac{1}{5}} \geq 2C_{inf}^{-1}$. For all $n \geq N$, with probability greater than $1 - 2e^{-Cn\rho_n}$, $\widehat{\gamma_n} \in \left[\underline{\gamma_n}, \gamma_n\right]$ and $\widehat{\rho_n} \in \left[\rho_n, \underline{\rho_n}\right]$. □

To prove Theorem 3, we work conditionnaly on the event $\mathcal{E}$. Note that in the model (1), the law of the remaining entries $(A_{i,j})_{(i,j)\notin\Omega}$ is independent of $\mathcal{E}$. Since on $\mathcal{E}$ both $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^*$ belong to the set $\left[\underline{\gamma_n}, \underline{\rho_n}\right]$, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_2^2 &= \sum_{(i,j)\in\Omega}\left(\widehat{\boldsymbol{\Theta}}_{ij} - \boldsymbol{\Theta}^*_{ij}\right)^2 + \sum_{(i,j)\notin\Omega}\left(\widehat{\boldsymbol{\Theta}}_{ij} - \boldsymbol{\Theta}^*_{ij}\right)^2 \\ &\leq n\underline{\rho_n}^2 + \sum_{(i,j)\notin\Omega}\left(\widehat{\boldsymbol{\Theta}}_{ij} - \boldsymbol{\Theta}^*_{ij}\right)^2. \end{aligned} \tag{26}$$

We adapt the proof of Theorem 1 to bound the second term. Let $\underline{\epsilon_n} \triangleq C\left(\underline{\rho_n}\big/\underline{\gamma_n}\right)^2\left(n\log(k) + k^2\right)$ where C is the same absolute constant as in Theorem 1, and let $\underline{\epsilon}^0 \triangleq \underline{\rho_n\epsilon_n}$. We start by considering the following two cases:

Case 1: $\sum_{(i,j)\notin\Omega}\left(\widehat{\boldsymbol{\Theta}}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right)^2 \leq \underline{\epsilon}^0$. Then, the statement of Theorem 3 follows from (26) and Lemma 19:

$$\begin{aligned}\sum_{(i,j)\notin\Omega}\left(\widehat{\boldsymbol{\Theta}}_{ij} - \boldsymbol{\Theta}^*_{ij}\right)^2 &\leq 2\sum_{(i,j)\notin\Omega}\left(\widehat{\boldsymbol{\Theta}}_{ij} - \widetilde{\boldsymbol{\Theta}}_{ij}\right)^2 + 2\sum_{(i,j)\notin\Omega}\left(\widetilde{\boldsymbol{\Theta}}_{ij} - \boldsymbol{\Theta}^*_{ij}\right)^2 \\ &\leq 2\underline{\rho_n\epsilon_n} + 16\underline{\rho_n}\mathcal{K}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) \\ &\leq C\log(n)\rho_n\left(\mathcal{K}(\boldsymbol{\Theta}^*, \widetilde{\boldsymbol{\Theta}}) + n\log(k) + k^2\right).\end{aligned}$$

Case 2: $\sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\widetilde{\mathbf{\Theta}}_{ij}\right)^2>\underline{\epsilon}^0$. Then $\widehat{\mathbf{\Theta}}$ belongs to the set

$$\underline{\mathcal{S}}=\left\{\mathbf{\Theta}\in\bigcup_{z\in\mathcal{Z}_{n,k}}\mathcal{T}_z:\ \sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\widetilde{\mathbf{\Theta}}_{ij}\right)^2>\underline{\epsilon}^0,\ \|\mathbf{\Theta}\|_\infty\leq\underline{\rho_n},\ \min_{i<j}\{\mathbf{\Theta}_{ij}\}\geq\underline{\gamma_n}\right\}.$$

As before, Lemma 19 implies

$$\begin{aligned}\sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\mathbf{\Theta}^*_{ij}\right)^2 &\leq 8\underline{\rho_n}\sum_{(i,j)\notin\Omega}\mathcal{K}(\mathbf{\Theta}^*_{ij},\widehat{\mathbf{\Theta}}_{ij})\\ &\leq 8\underline{\rho_n}\mathcal{K}(\mathbf{\Theta}^*,\widetilde{\mathbf{\Theta}})+8\underline{\rho_n}\sum_{(i,j)\notin\Omega}\left(\mathcal{K}(\mathbf{\Theta}^*_{ij},\widehat{\mathbf{\Theta}}_{ij})-\mathcal{K}(\mathbf{\Theta}^*_{ij},\widetilde{\mathbf{\Theta}}_{ij})\right).\end{aligned}$$

On the event $\mathcal{E}$, $\widetilde{\mathbf{\Theta}}$ belongs to the set of matrices on which the maximum likelihood estimator is defined, thus the definition of $\widehat{\mathbf{\Theta}}$ implies $\sum_{(i,j)\notin\Omega}\left(\mathcal{K}(\boldsymbol{A}_{ij},\widehat{\mathbf{\Theta}}_{ij})-\mathcal{K}(\boldsymbol{A}_{ij},\widetilde{\mathbf{\Theta}}_{ij})\right)\leq 0$ and

$$\sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\mathbf{\Theta}^*_{ij}\right)^2\leq 8\underline{\rho_n}\mathcal{K}(\mathbf{\Theta}^*,\widetilde{\mathbf{\Theta}})+8\underline{\rho_n}\sum_{(i,j)\notin\Omega}\left(\mathcal{K}(\mathbf{\Theta}^*_{ij},\widehat{\mathbf{\Theta}}_{ij})-\mathcal{K}(\boldsymbol{A}_{ij},\widehat{\mathbf{\Theta}}_{ij})-\left(\mathcal{K}(\mathbf{\Theta}^*_{ij},\widetilde{\mathbf{\Theta}}_{ij})-\mathcal{K}(\boldsymbol{A}_{ij},\widetilde{\mathbf{\Theta}}_{ij})\right)\right).$$

The proof of the following lemma follows the lines of the proof of Lemma 3, and we do not present it.

Lemma 18. *There exists a constant $C>0$ depending only on C_{inf} such that for all $\mathbf{\Theta}\in\underline{\mathcal{S}}$ simultaneously we have*

$$\left|\sum_{(i,j)\notin\Omega}\left(\mathcal{K}(\mathbf{\Theta}^*_{ij},\mathbf{\Theta}_{ij})-\mathcal{K}(\boldsymbol{A}_{ij},\mathbf{\Theta}_{ij})-\left(\mathcal{K}(\mathbf{\Theta}^*_{ij},\widetilde{\mathbf{\Theta}}_{ij})-\mathcal{K}(\boldsymbol{A}_{ij},\widetilde{\mathbf{\Theta}}_{ij})\right)\right)\right|\leq\frac{1}{32\underline{\rho_n}}\sum_{(i,j)\notin\Omega}\left(\mathbf{\Theta}_{ij}-\widetilde{\mathbf{\Theta}}_{ij}\right)^2+\underline{\epsilon_n}$$

with probability at least $1-5\exp\left(-C\underline{\rho_n}\left(n\log(k)+k^2\right)\right)$.

This implies that on the event $\mathcal{E}$, with large probability,

$$\begin{aligned}\sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\mathbf{\Theta}^*_{ij}\right)^2 &\leq 8\underline{\rho_n}\mathcal{K}(\mathbf{\Theta}^*,\widetilde{\mathbf{\Theta}})+\frac{1}{4}\sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\widetilde{\mathbf{\Theta}}_{ij}\right)^2+8\underline{\epsilon_n\rho_n}\\ &\leq 8\underline{\rho_n}\mathcal{K}(\mathbf{\Theta}^*,\widetilde{\mathbf{\Theta}})+\frac{1}{2}\sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\mathbf{\Theta}^*_{ij}\right)^2+\frac{1}{2}\sum_{(i,j)\notin\Omega}\left(\mathbf{\Theta}^*_{ij}-\widetilde{\mathbf{\Theta}}_{ij}\right)^2+8\underline{\epsilon_n\rho_n}\\ \frac{1}{2}\sum_{(i,j)\notin\Omega}\left(\widehat{\mathbf{\Theta}}_{ij}-\mathbf{\Theta}^*_{ij}\right)^2 &\leq \left(8\underline{\rho_n}+4\underline{\rho_n}\right)\mathcal{K}(\mathbf{\Theta}^*,\widetilde{\mathbf{\Theta}})+8\underline{\epsilon_n\rho_n}.\end{aligned}\tag{27}$$

Using (26) and (27), we have shown that for $n\geq N$ and some constants $C,C'>0$ depending only on C_{inf}, with probability at least $1-5\exp\left(-C\underline{\rho_n}\left(n\log(k)+k^2\right)\right)-2\exp\left(-Cn\rho_n\right)$,

$$\left\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right\|_2^2\leq C\left(\underline{\rho_n^2}n+\underline{\rho_n}\mathcal{K}(\mathbf{\Theta}^*,\widetilde{\mathbf{\Theta}})+\underline{\rho_n\epsilon_n}\right).$$

We conclude the proof of Theorem 3 by noticing that $n\underline{\rho_n}^2\leq\underline{\rho_n\epsilon_n}$ and using that $\underline{\rho_n}=C\log(n)^{\frac{1}{5}}\rho_n$.

4.7 Proof of Proposition 1

By definition,

$$\begin{aligned}\mathcal{K}\left(\mathbf{\Theta}^*,\mathbf{\Theta}^{bc}\right) &= \sum_{i<j}\left(\mathbf{\Theta}^*_{ij}\log\left(\frac{\mathbf{\Theta}^*_{ij}}{\mathbf{\Theta}^{bc}_{ij}}\right)+\left(1-\mathbf{\Theta}^*_{ij}\right)\log\left(\frac{1-\mathbf{\Theta}^*_{ij}}{1-\mathbf{\Theta}^{bc}_{ij}}\right)\right)\\ &\leq \sum_{i<j}\left(\mathbf{\Theta}^*_{ij}\frac{\mathbf{\Theta}^*_{ij}-\mathbf{\Theta}^{bc}_{ij}}{\mathbf{\Theta}^{bc}_{ij}}+\left(1-\mathbf{\Theta}^*_{ij}\right)\frac{\mathbf{\Theta}^{bc}_{ij}-\mathbf{\Theta}^*_{ij}}{1-\mathbf{\Theta}^{bc}_{ij}}\right)\\ &= \sum_{i<j}\frac{\left(\mathbf{\Theta}^{bc}_{ij}-\mathbf{\Theta}^*_{ij}\right)^2}{\left(1-\mathbf{\Theta}^{bc}_{ij}\right)\mathbf{\Theta}^{bc}_{ij}}\end{aligned}$$

where the second line follows from the fact that for all $x > 0$, $\log(x) \leq x - 1$. Since for all $1 \leq i < j \leq n$, $\mathbf{\Theta}^{bc}_{ij}$ and $\mathbf{\Theta}^{*}_{ij}$ belong to $[C_{inf}\rho_n, \rho_n]$, this yields

$$\mathcal{K}\left(\mathbf{\Theta}^{*}, \mathbf{\Theta}^{bc}\right) \quad \leq \quad \sum_{i<j} \frac{\left(\mathbf{\Theta}^{bc}_{ij} - \mathbf{\Theta}^{*}_{ij}\right)^2}{(1-\rho_n)\, C_{inf}\rho_n}.$$

Now, recall that $\mathbf{\Theta}^{*}_{ij} = \rho_n W(\zeta_i, \zeta_j)$ and define $z^* : [n] \to [k]$ by $z^*(i) = \sum_{1 \leq a \leq k} a \mathbb{1}\left\{\zeta_i \in \left[\frac{a-1}{k}, \frac{a}{k}\right)\right\}$ for all $i \in [n]$. Moreover, define $\mathbf{\Theta}^{bc}_{ij} = \rho_n W\left(\frac{z^*(i)}{k}, \frac{z^*(j)}{k}\right)$. Note that by definition of z^*, for all i, $\left|\zeta_i - \frac{z^*(i)}{k}\right| \leq \frac{1}{k}$. Thus

$$\begin{aligned}
\mathcal{K}\left(\mathbf{\Theta}^{*}, \mathbf{\Theta}^{bc}\right) \quad &\leq \quad \frac{\rho_n}{C_{inf}(1-\rho_n)} \sum_{i<j} \left(W(\zeta_i, \zeta_j) - W\left(\frac{z^*(i)}{k}, \frac{z^*(j)}{k}\right)\right)^2 \\
&\leq \quad \frac{4\rho_n M^2}{C_{inf}(1-\rho_n)} \sum_{i<j} \left(\frac{1}{k}\right)^{2(\alpha \wedge 1)}
\end{aligned}$$

where the last equation follows from (12).

4.8 Technical lemmas

Lemma 19. *For all $\mathbf{\Theta}$, $\mathbf{\Theta}' \in \mathbb{R}^{n \times n}$ and $\mathbf{\Pi} \in [0,1]^{n \times n}_{sym}$,*

$$\|\mathbf{\Theta} - \mathbf{\Theta}'\|^2_{2,\mathbf{\Pi}} \leq 8\left(\|\mathbf{\Theta}\|_\infty \vee \|\mathbf{\Theta}'\|_\infty\right) \mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}, \mathbf{\Theta}').$$

Proof. By definition of Bernoulli Kullback-Leibler divergence for any $0 < q, q' < 1$ we have that

$$\begin{aligned}
\mathcal{K}(q, q') &= q \log\left(\frac{q}{q'}\right) + (1-q)\log\left(\frac{1-q}{1-q'}\right) \geq \left(\sqrt{q} - \sqrt{q'}\right)^2 + \left(\sqrt{1-q} - \sqrt{1-q'}\right)^2 \\
&\geq \frac{1}{2}\left[\left(\sqrt{q} - \sqrt{q'}\right) - \left(\sqrt{1-q} - \sqrt{1-q'}\right)\right]^2.
\end{aligned}$$

Using Taylor's Theorem for some η between q and q' we get

$$\begin{aligned}
\mathcal{K}(q, q') &\geq \frac{1}{2}\left[\frac{1}{2\sqrt{\eta}}(q - q') + \frac{1}{2\sqrt{1-\eta}}(q - q')\right]^2 = \frac{(q-q')^2}{8}\left[\frac{1}{\sqrt{\eta}} + \frac{1}{\sqrt{1-\eta}}\right]^2 \\
&= \frac{(q-q')^2}{8}\left[\frac{1}{\eta} + \frac{1}{1-\eta}\right] = \frac{(q-q')^2}{8}\frac{1}{\eta(1-\eta)} \geq \frac{(q-q')^2}{8(q \vee q')}.
\end{aligned} \tag{28}$$

Now Lemma 19 follows from (28) and

$$\mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}, \mathbf{\Theta}') = \sum_{i<j} \mathbf{\Pi}_{ij} \mathcal{K}(\mathbf{\Theta}_{ij}, \mathbf{\Theta}'_{ij}).$$

□

Lemma 20. *Let $\widetilde{\mathbf{\Theta}}^s$ and n_s be defined as in (8), and assume that $\gamma_n \leq \frac{1}{2}$, then*

$$\mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}^s) - \mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}) \leq 2\gamma_n n_s.$$

Proof.

$$\begin{aligned}
\mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*, \widetilde{\mathbf{\Theta}}^s) - \mathcal{K}_{\mathbf{\Pi}}(\mathbf{\Theta}^*_{ij}, \widetilde{\mathbf{\Theta}}_{ij}) \quad &= \quad \sum_{i<j} \mathbf{\Pi}_{ij}\left(\mathbf{\Theta}^*_{ij} \log\left(\frac{\widetilde{\mathbf{\Theta}}_{ij}}{\widetilde{\mathbf{\Theta}}^s_{ij}}\right) + (1 - \mathbf{\Theta}^*_{ij})\log\left(\frac{1-\widetilde{\mathbf{\Theta}}_{ij}}{1-\widetilde{\mathbf{\Theta}}^s_{ij}}\right)\right) \\
&= \quad \sum_{i<j} \mathbf{\Pi}_{ij} \mathbb{1}\left\{\widetilde{\mathbf{\Theta}}_{ij} < \gamma_n\right\}\left(\mathbf{\Theta}^*_{ij} \log\left(\frac{\widetilde{\mathbf{\Theta}}_{ij}}{\gamma_n}\right) + (1 - \mathbf{\Theta}^*_{ij})\log\left(\frac{1-\widetilde{\mathbf{\Theta}}_{ij}}{1-\gamma_n}\right)\right) \\
&\leq \quad \sum_{i<j} \mathbf{\Pi}_{ij} \mathbb{1}\left\{\widetilde{\mathbf{\Theta}}_{ij} < \gamma_n\right\}(1 - \mathbf{\Theta}^*_{ij})\log\left(1 + \frac{\gamma_n - \widetilde{\mathbf{\Theta}}_{ij}}{1-\gamma_n}\right) \\
&\leq \quad \sum_{i<j} \mathbf{\Pi}_{ij} \mathbb{1}\left\{\widetilde{\mathbf{\Theta}}_{ij} < \gamma_n\right\}\frac{\gamma_n - \widetilde{\mathbf{\Theta}}_{ij}}{1-\gamma_n} \leq \sum_{i<j} \mathbf{\Pi}_{ij} \mathbb{1}\left\{\widetilde{\mathbf{\Theta}}_{ij} < \gamma_n\right\} 2\gamma_n \leq 2n_s\gamma_n.
\end{aligned}$$

□

References

[1] E. Abbe and S. Sandon. Community Detection in General Stochastic Block models: Fundamental Limits and Efficient Algorithms for Recovery. 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, 670-688, 2015.

[2] N. Agarwal and A. S. Bandeira. Multisection in the Stochastic Block Model Using Semidefinite Programming. Compressed Sensing and its Applications: Second International MATHEON Conference 2015, Springer International Publishing, 125–162, 2017.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. Reviews of Modern Physics, 74:47–97, 2002.

[4] A. A. Amini, A. Chen, P. J. Bickel and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. The Annals of Statistics, 41(4):2097–2122, 2013.

[5] A. A. Amini and E. Levina On semidefinite relaxations for the block model. The Annals of Statistics, 46(1):149–179, 2018.

[6] A. S. Bandeira. Random Laplacian Matrices and Convex Relaxations. Foundations of Computational Mathematics, 18(2):345–379, 2018.

[7] O. Benyahia, C. Largeron, and B. Jeudy. Community detection in dynamic graphs with missing edges. 2017 11th International Conference on Research Challenges in Information Science (RCIS), 372-381, 2017.

[8] P. J. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. The Annals of Statistics, 41(4):1922–1943, 2013.

[9] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. Proceedings of the National Academy of Sciences, 106(50):21068–21073, 2009.

[10] K. Bleakley, G. Biau, and J.-P. Vert. Supervised reconstruction of biological networks with local models. Bioinformatics, 23(13):157-165, 2007.

[11] C. Bordenave, M. Lelarge, and L. Massoulié. Nonbacktracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs. The Annals of Probability, 46(1):1–71, 2018.

[12] E. J. Candès and B. Recht. Exact Matrix Completion via Convex Optimization. Foundations of Computational Mathematics, 9(6):717, 2009.

[13] A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electronic Journal of Statistics, 6:1847–1899, 2012.

[14] S. Chatterjee. Matrix estimation by Universal Singular Value Thresholding. The Annals of Statistics, 43:177–214, 2015.

[15] K. Chen and J. Lei. Network Cross-Validation for Determining the Number of Communities in Network Data. Journal of the American Statistical Association, 113:(521)241-251, 2018.

[16] A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. Nature, 453:8-101, 2008.

[17] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. Information and Inference. A Journal of the IMA, 3(3):189–223, 2014.

[18] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Physical review. E, Statistical, nonlinear, and soft matter physics, 84, 2011.

[19] C. Gao,Y. Lu, Z. Ma, and H. H. Zhou. Optimal Estimation and Completion of Matrices with Biclustering Structures. Journal of Machine Learning Research, 17(1):5602–5630, 2016.

[20] C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. The Annals of Statistics, 43(6):2624–2652, 2015.

[21] E. Giné and R. Nickl. Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2015.

[22] C. Giraud and N. Verzelen. Partial recovery bounds for clustering with the relaxed K-means. CoRR, 2018.

[23] R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Science, 106:22073–22078, 2009.

[24] L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 11(9):1074-1085, 1992.

[25] B. Hajek, Y. Wu, and J. Xu. Achieving Exact Cluster Recovery Threshold via Semidefinite Programming: Extensions. IEEE Transactions on Information Theory, 62(10):5918–59373, 2016.

[26] M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. The Annals of Applied Statistics, 4(1):5–25, 2010.

[27] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. Bernoulli, 20(1):282–303, 2014.

[28] O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. The Annals of Statistics, 45, 2017.

[29] O. Klopp, J. Lafond, E. Moulines, and J. Salmon. Adaptive Multinomial Matrix Completion. Electronic Journal of Statistics, 9, 2014.

[30] O. Klopp and N. Verzelen. Optimal graphon estimation in cut distance. Probability Theory and Related Fields, 2018.

[31] V. Koltchinskii , K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics, 39(5):2302–2329, 2011.

[32] G. Kossinets. Effects of missing data in social networks. Social Networks, 28(3):247 - 268, 2006.

[33] M.Kshirsagar, J. G. Carbonell, and J. Klein-Seetharaman. Techniques to cope with missing data in hostpathogen protein interaction prediction. Bioinformatics, 2012.

[34] P. Latouche, E. Birmel, and C. Ambroise. Overlapping Stochastic Block Models with Application to the French Political Blogosphere. Annals of Applied Statistics, 2009.

[35] J.-B. Leger, C. Vacher, and J.-J. Daudin. Detection of structurally homogeneous subsets in graphs. Statistics and Computing, 24(5):675–692, 2014.

[36] J. Lei. A goodness-of-fit test for stochastic block models. The Annals of Statistics, 44(1):401–424, 2016.

[37] L. Lovász. Large networks and graph limits. American Mathematical Society Colloquium Publications, American Mathematical Society, 2012.

[38] L. Lü and T. Zhou. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 390(6):1150 - 1170, 2011.

[39] L. Massoulié. Community detection thresholds and the weak Ramanujan property Proceedings of the Annual ACM Symposium on Theory of Computing, 2013.

[40] C. Matias, and S. Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. ESAIM: Proc., 47:55–74, 2014.

[41] F. McSherry. Spectral Partitioning of Random Graphs. Proceedings of the 42Nd IEEE Symposium on Foundations of Computer Science, 2001.

[42] E. Mossel, J. Neeman, and A. Sly. Consistency Thresholds for the Planted Bisection Model. Electronic Journal of Probability, 21, 2016.

[43] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics, 39:(2)069–1097, 2011.

[44] M. E. J. Newman. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23):8577–8582, 2006.

[45] S. C. Olhede and P. J. Wolfe. Network histograms and universality of blockmodel approximation. Proceedings of the National Academy of Sciences, 1111(41):714722–14727, 014.

[46] F. Picard, V. Miele, L. Cottret, J.-J. Daudin, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. BMC Bioinformatics, 2009.

[47] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. The Annals of Statistics, 139(4):1878–1915, 2011.

[48] T. Tabouy, P. Barbillon, and J. Chiquet. Variational Inference for Stochastic Block Models from Sampled Data. ArXiv e-prints, 2017.

[49] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Cambridge University Press, 2012.

[50] P. Wang, B. Xu, Y. Wu, and X. Zhou. Link prediction in social networks: the state-of-the-art. Science China Information Sciences, 58(1):1–38, 2015.

[51] Y. X. R. Wang and P. J. Bickel. Likelihood-based model selection for stochastic block models. The Annals of Statistics, 45(2):500–528, 017.

[52] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences, Cambridge University Press, 1994.

[53] J. Xu. Rates of Convergence of Spectral Methods for Graphon Estimation. Proceedings of the 35th International Conference on Machine Learning, 80:5433–5442, 2018.

[54] Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. IBioinformatics, 2004.

[55] B. Yan and S. Gregory. Finding missing edges in networks based on their community structure. Physical review. E, Statistical, nonlinear, and soft matter physics, 2012.

[56] Y. Zhang, E. Levina, and J. Zhu. Estimating network edge probabilities by neighbourhood smoothing. Biometrika, 104(4):771-783, 2017.

[57] Y. Zhao, Y.-J. Wu, E. Levina, and J. Zhu. Link Prediction for Partially Observed Networks. Journal of Computational and Graphical Statistics, 26(3):725-733, 2017.