Statistical and Computational Rates in Graph Logistic Regression

Nicolai Baldin
University of Cambridge

Quentin Berthet
Google Research, Brain Team

Abstract

We consider the problem of graph logistic regression, based on partial observation of a large network, and on side information associated to its vertices. The generative model is formulated as a matrix logistic regression. The performance of the model is analyzed in a high-dimensional regime under a structural assumption. The optimal statistical rates are derived, and an estimator based on penalized maximum likelihood is shown to attain it. The algorithmic aspects of this problem are also studied, and optimal rates under computational constraints are derived, and shown to differ from the information-theoretic rates - under a complexity assumption.

1 Introduction

In the field of network analysis, the task of *link prediction* consists in predicting the presence or absence of edges in a large graph, based on the observations of some of its edges, and on side information. Network analysis has become a growing inspiration for statistical problems. Indeed, one of the main characteristics of datasets in the modern scientific landscape is not only their growing size, but also their increasing complexity. Most phenomena now studied in the natural and social sciences concern not only isolated and independent variables, but also their interactions and connections.

Most statistical problems based on graphs are unsupervised: the graph itself is the sole data, there is no side information, and the objective is to recover an unknown structure in the generative model. Examples include the planted clique problem (Kučera, 1995; Alon and Sudakov, 1998), the stochastic block model Holland et al. (1983)—see Abbe (2017) for a recent survey of a very active line of work Decelle et al. (2011); Mossel et al. (2013); Massoulié (2014); Mossel et al. (2015); Abbe and Sandon (2015); Banks et al. (2016), the Ising blockmodel Berthet et al. (2016), random geometric graphs – see Penrose (2003) for an introduction and Devroye et al. (2011); Bubeck et al. (2014) for recent developments in statistics, or metric-based learning Chen et al. (2009); Bellet et al. (2014) and ordinal embeddings Jain et al. (2016).

In supervised regression problems on the other hand, the focus is on understanding a fundamental mechanism, formalized as the link between two variables. The objective is to learn how an explanatory variable X allows to predict a response Y, i.e. to find the unknown function f that best approximates the relationship $Y \approx f(X)$. This statistical framework is often applied to the observation of a phenomenon measured by Y (e.g. of a natural or social nature), given known information X: the principle is to understand said phenomenon, to explain the relationship between the variables by estimating the function f (Holland and Leinhardt, 1981; Hoff et al., 2002).

We follow this approach here: our goal is to learn how known characteristics of each agent (represented by a node) in the network induce a greater or smaller chance of connection, to understand the mechanism of formation of the graph. We propose a model for supervised inference on graphs. For each vertex, we are given side information: a vector of observations $X \in \mathbf{R}^d$. Given observations X_i, X_j about nodes i and j of a network, we aim to understand how these two explanatory variables are related to the probability of connection between the two corresponding vertices, such that $\mathbf{P}(Y_{(i,j)} = 1) = f(X_i, X_j)$, by estimating f within a high-dimensional class based on logistic regression. Besides this high-dimensional parametric modelling, various fully non-parametric statistical frameworks were exploited in the literature, see, for example, Wolfe and Olhede (2013); Gao et al. (2015), for graphon estimation, Papa et al. (2016); Biau and Bleakly (2008) for graph reconstruction and Bickel and Chen (2009) for modularity analysis.

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

Graph regression can be useful in any application where data can be gathered about the nodes of a network. With access to side information about each member of a social network, the objective is to understand the mechanisms of connection between members: shared interests, artistic tastes (Wasserman and Faust, 1994; Liben-Nowell and Kleinberg, 2003). This can also be applied to citation networks, or in the natural sciences to biological networks of interactions between molecules or proteins (Yu et al., 2008; Madeira and Oliveira, 2004). The key assumption in this model is that the network is a consequence of the information, but not necessarily based on similarity: it is possible to model more complex interactions, i.e. where opposites attract.

We therefore decide to tackle graph regression by modelling it as matrix logistic regression. We study a generative model for which $\mathbf{P}(Y_{(i,j)} = 1) = \sigma(X_i^\top \Theta_\star X_j)$, where σ is the sigmoid function, and $\Theta_\star$ is the unknown matrix to estimate. It is a simple way to model how the variables *interact*, by a quadratic *affinity* function and a sigmoid function. In order to model realistic situations with partial observations, we assume that $Y_{(i,j)}$ is only observed for a subset of all the pairs (i, j), denoted by Ω.

The matrix logistic regression model can be seen as a type of a high-dimensional logistic regression, using vectorization, and is naturally linked to trace regression models and graphons, which have received a significant amount of attention in the last years (Wolfe and Olhede, 2013; Klopp and Tsybakov, 2015; Gao et al., 2015; Zhang et al., 2015; Fan et al., 2016, 2017). The recent sparkle of interest in these models is partly driven by technical challenges in the analysis of statistical and computational performances in high-dimensional settings.

To convey the general idea of a simple dependency of $Y_{(i,j)}$ on X_i and X_j, we make structural assumptions on the rank and sparsity of $\Theta_\star$. This reflects that the affinity $X_i^\top \Theta_\star X_j$ is a function of the projections $u_\ell^\top X$ for the vectors X_i and X_j, for a small number of orthogonal vectors, that have themselves a small number of non-zero coefficients (sparsity assumption). In order to impose that the inverse problem is well-posed, we also make a restricted conditioning assumption on $\Theta_\star$, inspired by the restricted isometry property (RIP). These conditions are discussed in Section 2.

The classical techniques of likelihood maximization can lead to computationally intractable optimization problems. We show that in this problem as well as others this is a fundamental difficulty, not a weakness of one particular estimation technique; statistical and computational complexities are intertwined. Our findings are depicted in Figure 1.

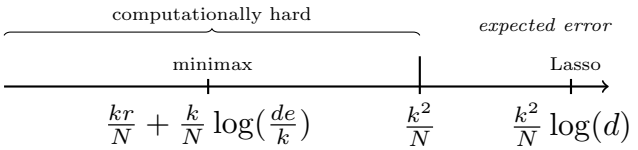


Figure 1: The computational and statistical boundaries for estimation and prediction in the matrix logistic regression model located on the real line. Here k denotes the sparsity of $\Theta_\star$ and r its rank, while N is the number of observed edges in the network.

Notation: For any positive integer n, we denote by $[n]$ the set $\{1, \dots, n\}$ and by $[[n]]$ the set of pairs of $[n]$, of cardinality $\binom{n}{2}$. We denote by $\mathbf{R}$ the set of real numbers and by $\mathbf{S}^d$ the set of real symmetric matrices of size d. For a matrix $A \in \mathbf{S}^d$, we denote by $\|A\|_F$ its Frobenius norm, defined by

$$\|A\|_F^2 = \sum_{i,j\in[d]} A_{ij}^2 \,.$$

We extend this definition for $B \in \mathbf{S}^n$ and any subset $\Omega \subseteq [[n]]$ to its semi-norm $\|B\|_{F,\Omega}$ defined by

$$\|B\|_{F,\Omega}^2 = \sum_{i,j\,:\,(i,j)\in\Omega} B_{ij}^2 \,.$$

The corresponding bilinear form playing the role of inner-product of two matrices $B_1, B_2 \in \mathbf{S}^n$ is denoted as $\langle\!\langle B_1, B_2 \rangle\!\rangle_{F,\Omega}$. For a matrix $B \in \mathbf{S}^n$, we also make use of the following matrix norms and pseudo-norms for $p, q \in [0, \infty)$, with $\|B\|_{p,q} = \big\|(\|B_{1*}\|_p \cdots \|B_{d*}\|_p)\big\|_q$, where B_{i*} denotes the ith row of B, and $\|B\|_\infty = \max_{(i,j)\in[d]} |B_{ij}|$. For asymptotic bounds, we shall write $f(x) \lesssim g(x)$ if $f(x)$ is bounded by a constant multiple of $g(x)$.

2 Problem description

2.1 Generative model

For a set of vertices $V = [n]$ and explanatory variables $X_i \in \mathbf{R}^d$ associated to each $i \in V$, a random graph $G = (V, E)$ is generated by the following model. For all $i, j \in V$, variables $X_i, X_j \in \mathbf{R}^d$ and an unknown matrix $\Theta_\star \in \mathbf{S}_d$, an edge connects the two vertices i and j independently of the others according to the distribution

$$\mathbf{P}\big((i,j) \in E\big) = \sigma(X_i^\top \Theta_\star X_j) = \frac{1}{1 + \exp(-X_i^\top \Theta_\star X_j)} \,. \tag{2.1}$$

Here we denote by σ the *sigmoid*, or *logistic* function.

Definition 1. *We denote by* $\pi_{ij} : \mathbf{S}_d \to [0,1]$ *the function mapping a matrix* $\Theta \in \mathbf{S}_d$ *to the probability in* (2.1). *Let* $\Sigma \in \mathbf{S}_n$ *with* $\Sigma_{ij} = X_i^\top \Theta X_j$ *denote the so-called affinity matrix. In particular, we then have* $\pi_{ij}(\Theta) = \sigma(\Sigma_{ij})$.

Our observation consists of the explanatory variables X_i and of the observation of a subset of the graph. Formally, for a subset $\Omega \subseteq [[n]]$, we observe an adjacency vector Y indexed by Ω that satisfies, for all $(i,j) \in \Omega$, $Y_{(i,j)} = 1$ if and only if $(i,j) \in E$ (and 0 otherwise). We thus have

$$Y_{(i,j)} \sim \text{Bernoulli}\big(\pi_{ij}(\Theta_\star)\big)\,, \quad (i,j) \in \Omega\,. \tag{2.2}$$

The joint data distribution is denoted by $\mathbf{P}_{\Theta_\star}$ and is thus completely specified by $\pi_{ij}(\Theta_\star)$, $(i,j) \in \Omega$. For ease of notation, we write $N = |\Omega|$, representing the *effective sample size*, which is the size of the set of observed edges. If all edges are observed, it equals $n(n-1)/2$, but it can be significantly smaller if only a subset of edges is observed. Our objective is to estimate the parameter matrix $\Theta_\star$, based on the observations $Y \in \mathbf{R}^N$ and on known explanatory variables $\mathbb{X} \in \mathbf{R}^{d\times n}$.

This problem can be reformulated as a classical logistic regression problem. Indeed, writing $\mathbf{vec}(A) \in \mathbf{R}^{d^2}$ for the *vectorized* form of a matrix $A \in \mathbf{S}^d$, we have that

$$X_i^\top \Theta_\star X_j = \mathbf{Tr}(X_j X_i^\top \Theta_\star) = \langle \mathbf{vec}(X_j X_i^\top), \mathbf{vec}(\Theta_\star)\rangle\,. \tag{2.3}$$

The vector of observation $Y \in \mathbf{R}^N$ therefore follows a logistic distribution with explanatory design matrix $\mathbb{D}_\Omega \in \mathbf{R}^{N\times d^2}$ such that $\mathbb{D}_{\Omega\,(i,j)} = \mathbf{vec}(X_j X_i^\top)$ and predictor $\mathbf{vec}(\Theta_\star) \in \mathbf{R}^{d^2}$. We focus on the matrix formulation of this problem, and consider directly the matrix form of *graph logistic regression* in order to simplify the notation of the explanatory variables and our model assumptions on $\Theta_\star$, that are specific to matrices.

2.2 Connection to other models

This model can be compared to other settings in the statistical and learning literature.

Generalised linear model. As discussed above in the remark to (2.3), this is an example of a logistic regression model. We focus in this work on the case where the matrix $\Theta_\star$ is block-sparse. The problem of sparse generalised linear models, and sparse logistic regression in particular has been extensively studied, see, e.g. (van de Geer, 2008; Bunea, 2008; Meier et al., 2008; Bach, 2010; Rigollet, 2012; Abramovich and Grinshtein, 2016) and references therein. Our work focuses on the more restricted case of block-sparse and low-rank parameter, establishing interesting statistical and computational phenomena in this setting.

Graphon model. The graphon model is a model of a random graph in which the explanatory variables associated with the vertices in the graph are unknown. It has recently become popular in the statistical community, see (Wolfe and Olhede, 2013; Klopp and Tsybakov, 2015; Gao et al., 2015; Zhang et al., 2015). Typically, an objective of statistical inference is a link function which belongs to either a parametric or non-parametric class of functions. Interestingly, the minimax lower bound for the classes of Hölder-continuous functions, obtained in (Gao et al., 2015), has not been attained by any polynomial-time algorithm.

Trace regression models. The modelling assumption (2.1) of the present paper is in fact very close to the trace regression model, as it follows from the representation (2.3). Thus, the block-sparsity and low-rank structures are preserved and can well be studied by the means of techniques developed for the trace regression. We refer to (Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Rohde and Tsybakov, 2011; Fan et al., 2016) for recent developments in the linear trace regression model, and (Fan et al., 2017) for the generalised trace regression model. However, computational lower bounds have not been studied either and many existing minimax optimal estimators cannot to be computed in polynomial time.

Metric learning. In the task of metric learning, observations depend on an unknown geometric representation $V_1, \dots, V_n$ of the variables in a Euclidean space of low dimension. The goal is to estimate this representation (up to a rigid transformation), based on noisy observations of $\langle V_i, V_j\rangle$ in the form of random evaluations of similarity. Formally, our framework also recovers the task of metric learning by taking $X_i = e_i$ and $\Theta_\star$ an unknown semidefinite positive matrix of small rank (here $V^\top V$), since

$$\langle V_i, V_j\rangle = \langle V e_i, V e_j\rangle = e_i^\top\, V^\top V\, e_j\,.$$

We refer to (Chen et al., 2009; Bellet et al., 2014) and references therein for a comprehensive survey of metric learning methods.

2.3 Parameter space

The unknown predictor matrix $\Theta_\star$ describes the relationship between the observed features X_i and the probabilities of connection $\pi_{ij}(\Theta_\star) = \sigma(X_i^\top \Theta_\star X_j)$ following Definition 1. We focus on the high-dimensional setting where $d^2 \gg N$: the number of features for each vertex in the graph, and number of free parameters, is much greater than the total number of observations. In order to counter the curse of dimen-

sionality, we make the assumption that the function $(X_i, X_j) \mapsto \pi_{ij}$ depends only on a small subset S of size k of all the coefficients of the explanatory variables. This translates to a *block-sparsity* assumption on $\Theta_\star$: the coefficients $\Theta_{\star\, ij}$ are only nonzero for i and j in S. Furthermore, we assume that the rank of the matrix $\Theta_\star$ can be smaller than the size of the block. Formally, we define the following parameter spaces

$$\mathcal{P}_{k,r}(M) = \Big\{\Theta \in \mathbf{S}^d \,:\, \|\Theta\|_{1,1} < M \;,\; \|\Theta\|_{0,0} \leq k \;, \\ \text{and } \mathbf{rank}(\Theta) \leq r\Big\},$$

for the coefficient-wise ℓ_1 norm $\|\cdot\|_{1,1}$ on $\mathbf{S}^d$ and integers $k, r \in [d]$. We also denote $\mathcal{P}(M) = \mathcal{P}_{d,d}(M)$ for convenience.

Remark 2. *The bounds on block-sparsity and rank in our parameter space are structural bounds: we consider the case where the matrix $\Theta_\star$ can be concisely described in terms of the number of parameters. This is motivated by considering the spectral decomposition of the real symmetric matrix $\Theta_\star$ as*

$$\Theta_\star = \sum_{\ell=1}^{r} \lambda_\ell u_\ell u_\ell^\top \,.$$

The affinity $\Sigma_{ij} = X_i^\top \Theta_\star X_j$ between vertices i and j is therefore only a function of the projections of X_i and X_j along the axes u_ℓ, i.e.

$$\Sigma_{ij} = X_i^\top \Theta_\star X_j = \sum_{\ell=1}^{r} \lambda_\ell (u_\ell^\top X_i)(u_\ell^\top X_j) \,.$$

Assuming that there are only a few of these directions u_ℓ with non-zero impact on the affinity motivates the low-rank assumption, while assuming that there are only few relevant coefficients of X_i, X_j that influence the affinity corresponds to a sparsity assumption on the u_ℓ, or block sparsity of $\Theta_\star$. The effect of these projections on the affinity is weighted by the λ_ℓ. By allowing for negative eigenvalues, we allow our model to go beyond a geometric description, where close or similar Xs are more likely to be connected. This can be used to model interactions where opposites attract.

The assumption of simultaneously sparse and low-rank matrices arises naturally in many applications in statistics and machine learning and has attracted considerable recent attention. Various regularisation techniques have been developed for estimation, variable and rank selection in multivariate regression problems, see, e.g. (Bunea et al., 2012; Richard et al., 2012) and the references therein.

2.4 Explanatory variables

As mentioned above, this problem is different from tasks such as metric learning, where no side information is present, and "information-less" covariates $X_i = e_i$ are used to estimate an unknown geometric representation of the variables in a Euclidean space of low dimension. Here X_i are seen as covariates, allowing us to infer from the observation on the graph the predictor variable $\Theta_\star$. For this task to be even possible in a high-dimensional setting, we settle the identifiability issue by making the following variant of a classical assumption on $\mathbb{X} \in \mathbf{R}^{d\times n}$.

Definition 3 (Block isometry property)**.** *For a matrix $\mathbb{X} \in \mathbf{R}^{d\times n}$ and an integer $s \in [d]$, we define $\Delta_{\Omega,s}(\mathbb{X}) \in (0,1)$ as the smallest positive real such that*

$$\frac{\|\mathbb{X}^\top B\, \mathbb{X}\|_{F,\Omega}^2}{N} \in \big[\big(1 \pm \Delta_{\Omega,s}(\mathbb{X})\big)\|B\|_F^2\big] \,,$$

for all matrices $B \in \mathbf{S}^d$ that satisfy the block-sparsity assumption $\|B\|_{0,0} \leq s$.

Definition 4 (Restriced isometry properties)**.** *For a matrix $A \in \mathbf{R}^{n\times p}$ and an integer $s \in [p]$, $\delta_s(A) \in (0,1)$ is the smallest positive real such that*

$$n\big(1 - \delta_s(A)\big)\|v\|_2^2 \leq \|Av\|_2^2 \leq n\big(1 + \delta_s(A)\big)\|v\|_2^2 \,,$$

for all s-sparse vectors, i.e. satisfying $\|v\|_0 \leq s$.

When $p = d^2$ is a square, we define $\delta_{\mathcal{B},s}(A)$ as the smallest positive real such that

$$n\big(1 - \delta_{\mathcal{B},s}(A)\big)\|v\|_2^2 \leq \|Av\|_2^2 \leq n\big(1 + \delta_{\mathcal{B},s}(A)\big)\|v\|_2^2 \,,$$

for all vectors such that $v = \boldsymbol{vec}(B)$, where B satisfies the block-sparsity assumption $\|B\|_{0,0} \leq s$.

The first definition is due to (Candes and Tao, 2005), with restriction to sparse vectors. It can be extended in general, as here, to other types of restrictions, see e.g. (Traonmilin and Gribonval, 2015). Since the restriction on the vectors in the second definition (s-by-s block-sparsity) is more restricting than in the first one (sparsity), $\delta_{\mathcal{B},s}$ is smaller than δ_{s^2}. These different measures of restricted isometry are related, as shown in the following lemma

Lemma 5. *For a matrix $\mathbb{X} \in \mathbf{R}^{d\times n}$, let $\mathbb{D}_\Omega \in \mathbf{R}^{N\times d^2}$ be defined row-wise by $\mathbb{D}_{\Omega\,(i,j)} = \boldsymbol{vec}(X_j X_i^\top)$ for all $(i,j) \in \Omega$. It holds that*

$$\Delta_{\Omega,s}(\mathbb{X}) = \delta_{\mathcal{B},s}(\mathbb{D}_\Omega) \,.$$

Proof. This is a direct consequence of the definition of $\mathbb{D}_\Omega$, which yields $\|\mathbb{X}^\top B\, \mathbb{X}\|_{F,\Omega}^2 = \|\mathbb{D}_\Omega\, \mathbf{vec}(B)\|_2^2$, and $\|\mathbf{vec}(B)\|_2^2 = \|B\|_F^2$. □

The assumptions above guarantee that the matrix $\Theta_\star$ can be recovered from observations of the affinities,

settling the well-posedness of this part of the inverse problem. However, we do not directly observe these affinities, but their image through the sigmoid function. We must therefore further impose the following assumption on the design matrix $\mathbb{X}$ that yields constraints on the probabilities π_{ij} and in essence governs the identifiability of $\Theta_\star$.

Assumption 6. *There exists a constant $M > 0$ such that for all Θ in the class $\mathcal{P}(M)$ we have $\max_{(i,j)\in\Omega} |X_i^\top \Theta X_j| < M$.*

In particular, under this assumption a constant

$$\mathcal{L}(M) := \sigma'(M) = \sigma(M)\big(1-\sigma(M)\big)\,, \tag{2.4}$$

is lower bounded away from zero, and we have

$$\inf_{\Theta\in\mathcal{P}(M)} \sigma'(X_i^\top \Theta X_j) \geq \mathcal{L}(M) > 0\,, \tag{2.5}$$

for all $(i,j)\in\Omega$. Assuming that $\mathcal{L}(M)$ always depends on the same M, we sometimes write simply $\mathcal{L}$.

Remark 7. *Assumption 6 is necessary for the identifiability of $\Theta_\star$: if $X_i^\top \Theta_\star X_j$ can be arbitrarily large in magnitude, $\pi_{ij} = \sigma(X_i^\top \Theta_\star X_j)$ can be arbitrarily close to 0 or 1. Since our observations only depend on $\Theta_\star$ through its image π_{ij}, this could lead to a very large estimation error on $\Theta_\star$ even with a small estimation error on the π_{ij}.*

Remark 8. *This assumption has already appeared in the literature on high-dimensional estimation, see (van de Geer, 2008; Abramovich and Grinshtein, 2016).*

Proposition 9. *The identifiability assumption $\max_{(i,j)\in\Omega} |X_i^\top \Theta X_j| < M$ is guaranteed for all $\Theta \in \mathcal{P}(M)$ and design matrices $\mathbb{X}$ satisfying either of the following*

- $\|X_j X_i^\top\|_\infty \leq 1$,
- $\|\Theta\|_F^2 < M_1/N$ *for some* $M_1 > 0$ *and the block isometry property.*

2.4.1 Random designs

For random designs, we require the block isometry property to hold with high probability. Then the results in this article carry over directly and thus we do not discuss it in full detail. It is well known that for sparse linear models with the dimension of a target vector $\bar{p}$ and the sparsity $\bar{k}$, the classical restricted isometry property holds for some classes of random matrices with i.i.d. entries including sub-Gaussian and Bernoulli matrices, see (Mendelson et al., 2008), provided that $\bar{n} \gtrsim \bar{k}\log(\bar{p}/\bar{k})$, and i.i.d. subexponential random matrices, see (Adamczak et al., 2011), provided that $\bar{n} \gtrsim \bar{k}\log^2(\bar{p}/\bar{k})$. In the same spirit, the design matrices with independent entries following sub-Gaussian, subexponential or Bernoulli distributions can be shown to satisfy the block isometry property, cf. (Wang et al., 2016a), provided that the number of observed edges in the network satisfies $N \gtrsim k^2\log^2(d/k)$ for sub-Gaussian and subexponential designs and $N \gtrsim k^2\log(d/k)$ for Bernoulli designs.

3 Matrix Logistic Regression

The log-likelihood for this problem is

$$\ell_Y(\Theta) = -\sum_{(i,j)\in\Omega} \xi(s_{(i,j)} X_i^\top \Theta X_j)\,,$$

where $s_{(i,j)} = 2Y_{(i,j)} - 1$ is a sign variable that depends on the observations Y and $\xi : x \mapsto \log(1+e^x)$ is a *softmax* function, convex on $\mathbf{R}$. As a consequence, the negative log-likelihood $-\ell_Y$ is a convex function of Θ. Denoting by ℓ the expectation $\mathbb{E}_{\Theta_\star}[\ell_Y]$, we recall the classical expressions for all $\Theta \in \mathbf{S}^d$

$$\ell(\Theta) = \ell(\Theta_\star) - \mathsf{KL}(\mathbf{P}_{\Theta_\star}, \mathbf{P}_\Theta)\,,$$

where we recall $\pi_{ij}(\Theta) = \sigma(X_i^\top \Theta X_j)$, and

$$\ell_Y(\Theta) = \ell(\Theta) + \langle\!\langle \nabla\zeta, \Theta \rangle\!\rangle_F\,,$$

where ζ is a stochastic component of the log-likelihood with constant gradient $\nabla\zeta \in \mathbf{R}^{d\times d}$ given by $\nabla\zeta = \sum_{(i,j)\in\Omega}(Y_{(i,j)} - \pi_{ij}(\Theta_\star))X_j X_i^\top$, which is a sum of independent centered random variables.

3.1 Penalized logistic loss

In a classical setting where d is fixed and N grows, the maximiser of ℓ_Y - the maximum likelihood estimator - is an accurate estimator of $\Theta_\star$, provided that it is possible to identify Θ from $\mathbf{P}_\Theta$ (i.e. if the X_i are well conditioned). We are here in a high-dimensional setting where $d^2 \gg N$, and this approach is not directly possible. Our parameter space indicates that the intrinsic dimension of our problem is truly much lower in terms of rank and block-sparsity. Our assumption on the conditioning of the X_i is tailored to this structural assumption. In the same spirit, we also modify our estimator in order to promote the selection of elements of low rank and block-sparsity. Following the ideas of (Birgé and Massart, 2007) and (Abramovich and Grinshtein, 2016), we define the following penalized maximum likelihood estimator

$$\hat{\Theta} \in \operatorname*{argmin}_{\Theta\in\mathcal{P}(M)} \Big\{ -\ell_Y(\Theta) + p(\Theta) \Big\}\,, \tag{3.1}$$

with a penalty $p(\Theta) = g(\mathrm{rank}(\Theta), \|\Theta\|_{0,0})$, for

$$g(R,K) = cKR + cK\log\big(\frac{de}{K}\big)\,, \tag{3.2}$$

where $c > 0$ is a universal constant and to be specified further. The proof of the following theorem is based on Dudley's integral argument combined with Bousquet's inequality and is deferred to the Appendix.

Theorem 10. *Assume the design matrix $\mathbb{X}$ satisfies $\max_{(i,j)\in\Omega} |X_i^\top \Theta_\star X_j| < M$ for some $M > 0$ and all $\Theta_\star$ in a given class, and the penalty term $p(\Theta)$ satisfies (3.2) with the constants $c \geq c_1/\mathcal{L}$, $c_1 > 1$, $\mathcal{L}$ given in (2.4). Then for the penalised MLE estimator $\hat{\Theta}$, the following non-asymptotic upper bound on the expectation of the Kullback-Leibler divergence between the measures $\mathbf{P}_{\Theta_\star}$ and $\mathbf{P}_{\hat{\Theta}}$ holds*

$$\sup_{\Theta_\star \in \mathcal{P}_{k,r}(M)} \frac{\mathbb{E}[\mathsf{KL}(\mathbf{P}_{\Theta_\star}, \mathbf{P}_{\hat{\Theta}})]}{N} \leq C_1 \frac{kr}{N} + C_1 \frac{k}{N} \log\big(\frac{de}{k}\big)\,, \tag{3.3}$$

where $C_1 > 3c$ is some universal constant for all $k = 1, ..., d$ and $r = 1, ..., k$.

Remark 11. *Random designs with i.i.d. entries following sub-Gaussian, Bernoulli and subexponential distributions discussed in Section 2.4.1 yield the same rate as well. It can formally be shown using standard conditioning arguments, see, e.g. (Nickl and van de Geer, 2013).*

Corollary 12. *Assume the design matrix $\mathbb{X}$ satisfies the block isometry property from Definition 3 and $\max_{(i,j)\in\Omega} |X_i^\top \Theta_\star X_j| < M$ for some $M > 0$ and all $\Theta_\star$ in a given class, and the penalty term $p(\Theta)$ is as in Theorem 10. Then for the penalised MLE estimator $\hat{\Theta}$, the following non-asymptotic upper bound on the rate of estimation holds*

$$\sup_{\Theta_\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\hat{\Theta} - \Theta_\star\|_F^2\big] \leq \frac{C_1}{\mathcal{L}(M)\big(1 - \Delta_{\Omega,2k}(\mathbb{X})\big)} \Big(\frac{kr}{N} + \frac{k}{N} \log\big(\frac{de}{k}\big)\Big)\,,$$

where $C_1 > 3c$ is some universal constant for all $k = 1, ..., d$ and $r = 1, ..., k$.

Let us define rank-constrained maximum likelihood estimators with bounded block size as

$$\hat{\Theta}_{k,r} \in \operatorname*{argmin}_{\Theta \in \mathcal{P}_{k,r}(M)} \{-\ell_Y(\Theta)\}\,.$$

It is intuitively clear that without imposing any regularisation on the likelihood function, the maximum likelihood approach selects the most complex model. In fact, the following result holds.

Theorem 13. *Assume the design matrix $\mathbb{X}$ satisfies the block isometry property from Definition 3 and $\max_{(i,j)\in\Omega} |X_i^\top \Theta_\star X_j| < M$ for some $M > 0$ and all $\Theta_\star$ in a given class. Then for the maximum likelihood estimator $\hat{\Theta}_{k,r}$, the following non-asymptotic upper bound on the rate of estimation holds*

$$\sup_{\Theta_\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\hat{\Theta}_{k,r} - \Theta_\star\|_F^2\big] \leq \frac{C_3}{\mathcal{L}(M)\big(1 - \Delta_{\Omega,2k}(\mathbb{X})\big)} \Big(\frac{kr}{N} + \frac{k}{N} \log\big(\frac{de}{k}\big)\Big)\,,$$

for all $k = 1, ..., d$ and $r = 1, ..., k$ and some constant $C_3 > 0$.

Remark 14. *The penalty (3.2) belongs to the class of the so-called minimal penalties, cf. (Birgé and Massart, 2007). In particular, a naive MLE approach with $p(\Theta) = 0$ in (3.1) yields a suboptimal estimator as it follows from Theorem 13.*

3.2 Prediction

In applications, as new users join the network, we are interested in predicting the probabilities of the links between them and the existing users. It is natural to measure the prediction error of an estimator $\hat{\Theta}$ by $\mathbb{E}\big[\sum_{(i,j)\in\Omega}(\pi_{ij}(\hat{\Theta}) - \pi_{ij}(\Theta_\star))^2\big]$ which is controlled according to the following result using the smoothness of the logistic function σ.

Theorem 15. *Under Assumption 6, we have the following rate for estimating the matrix of probabilities $\Sigma_\star = \mathbb{X}^\top \Theta_\star \mathbb{X} \in \mathbf{R}^{n\times n}$ with the estimator $\hat{\Sigma} = \mathbb{X}^\top \hat{\Theta} \mathbb{X} \in \mathbf{R}^{n\times n}$, for all $\Theta_\star \in \mathcal{P}_{k,r}(M)$:*

$$\frac{1}{2N}\mathbb{E}\big[\|\hat{\Sigma} - \Sigma_\star\|_{F,\Omega}^2\big] \leq \frac{C_1}{\mathcal{L}(M)} \Big(\frac{kr}{N} + \frac{k}{N} \log\big(\frac{de}{k}\big)\Big)\,,$$

with the constant C_1 from (3.3). The rate is minimax optimal, i.e. a minimax lower bound of the same asymptotic order holds for the prediction error of estimating the matrix of probabilities $\Sigma_\star = \mathbb{X}^\top \Theta_\star \mathbb{X} \in \mathbf{R}^{n\times n}$.

Remark 16. *Whilst Assumption 6 is needed for the estimation task, there are results on the prediction task in logistic regression showing that it can be unnecessary, see Bach (2010). Whilst the main focus of this paper is on estimation, it is a promising avenue for future research to weaken or even remove this assumption in the matrix logistic regression setting as well.*

3.3 Convex relaxation

In practice, computation of the estimator (3.1) is often infeasible. In essence, in order to compute it, we need to compare the likelihood functions over all possible subspaces $\mathcal{P}_{k,r}(M)$. Sophisticated step-wise model selection procedures allow to reduce the number of analysed models, see e.g. Section 3.3.2 in Hastie et al. (2001). However, they are not feasible in a high-dimensional setting either. We here consider the following estimator

$$\hat{\Theta}_{1,1} = \operatorname*{argmin}_{\Theta \in \mathbf{S}^d} \{-\ell_Y(\Theta) + \lambda \|\Theta\|_{1,1}\}\,, \tag{3.4}$$

with $\lambda > 0$ to be chosen further, which is equivalent to the logistic Lasso on $\mathbf{vec}(\Theta)$. Using standard arguments, cf. Example 1 in (van de Geer, 2008), combined with the block isometry property the following result immediately follows.

Theorem 17. *Assume the design matrix $\mathbb{X}$ satisfies the block isometry property from Definition 3 and $\max_{(i,j)\in\Omega} |X_i^\top \Theta_\star X_j| < M$ for some $M > 0$ and all $\Theta_\star$ in a given class. Then for $\lambda = C_4\sqrt{\log d}$, where $C_4 > 0$ is an appropriate universal constant, the estimator (3.4) satisfies for all $\Theta_\star \in \mathcal{P}_{k,r}(M)$*

$$\mathbb{E}\big[\|\hat{\Theta}_{1,1} - \Theta_\star\|_F^2\big] \leq \frac{C_5}{\mathcal{L}(M)\big(1 - \Delta_{\Omega,2k}(\mathbb{X})\big)} \frac{k^2}{N} \log d\,, \tag{3.5}$$

for all $k = 1, ..., d$ and $r = 1, ..., k$ and some universal constant $C_5 > 0$.

As one could expect the upper bound on the rate of estimation of our feasible estimator is independent of the true rank r. It is natural, when dealing with a low-rank and block-sparse objective matrix, to combine the nuclear penalty with either the $(2,1)$-norm penalty or the $(1,1)$-norm penalty of a matrix, cf. (Giraud, 2011; Koltchinskii et al., 2011; Bunea et al., 2012; Richard et al., 2012). In our setting, it can be easily shown that combining the $(1,1)$-norm penalty and the nuclear penalty yields the same rate of estimation $(k^2/N)\log d$. This appears to be inevitable in view of a computational lower bound, obtained in Section 4, which is independent of the rank as well. In particular, these findings partially answer a question posed in Section 6.4.4 in (Giraud, 2014). Results about prediction bounds are presented in the appendix.

3.4 Information-theoretic lower bounds

The following result demonstrates that the minimax lower bound on the rate of estimation matches the upper bound in Theorem 10 implying that the rate of estimation is minimax optimal.

Theorem 18. *Let the design matrix $\mathbb{X}$ satisfy the block isometry property. Then for estimating $\Theta_\star \in \mathcal{P}_{k,r}(M)$ in the matrix logistic regression model, the following lower bound on the rate of estimation holds*

$$\inf_{\hat{\Theta}} \sup_{\Theta_\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\hat{\Theta} - \Theta_\star\|_F^2\big] \geq \frac{C_2}{(1 + \Delta_{\Omega,2k}(\mathbb{X}))}\Big(\frac{kr}{N} + \frac{k}{N}\log\big(\frac{de}{k}\big)\Big)\,,$$

where the constant $C_2 > 0$ is independent of d, k, r and the infimum extends over all estimators $\hat{\Theta}$.

Remark 19. *The lower bounds of the same order hold for the expectation of the Kullback-Leibler divergence between the measures $\mathbf{P}_{\Theta_\star}$ and $\mathbf{P}_{\hat{\Theta}}$ and the prediction error of estimating the matrix of probabilities $\Sigma_\star = \mathbb{X}^\top \Theta_\star \mathbb{X} \in \mathbf{R}^{n\times n}$.*

4 Computational lower bounds

In this section, we investigate whether the lower bound in Theorem 18 can be achieved with an estimator computable in polynomial time. Recently, the gap between computational and statistical optimal performance has attracted a lot of attention in the statistical community. We refer to (Berthet and Rigollet, 2013a,b; Wang et al., 2016b; Gao et al., 2017; Zhang and Dong, 2017; Hajek et al., 2015; Chen, 2015; Ma and Wu, 2015; Chen and Xu, 2016) for computational lower bounds in high-dimensional statistics based on the planted clique problem (see below), (Berthet and Ellenberg, 2015) using hardness of learning parity with noise (Oymak et al., 2015) for denoising of sparse and low-rank matrices, (Agarwal, 2012) for computational trade-offs in statistical learning, as well as (Zhang et al., 2014) for worst-case lower bounds for sparse estimators in linear regression, as well as (Bruer et al., 2015; Chandrasekaran and Jordan, 2013) for another approach on computational trade-offs in statistical problems, as well as (Berthet and Chandrasekaran, 2016; Berthet and Perchet, 2017) on the management of these trade-offs. In order to establish a computational lower bound for the block-sparse matrix logistic regression, we exploit a reduction scheme from (Berthet and Rigollet, 2013a): we show that detecting a subspace of $\mathcal{P}_{k,r}(M)$ can be computationally as hard as solving the dense subgraph detection problem. An introduction to this problem, and the complexity assumption at hand, is presented in the appendix.

4.1 Reduction to the dense subgraph detection problem

We define in the appendix a class of parameter matrices representing instances of the planted dense subgraph problem, in the graph logistic regression model. In this problem, a conjecture is posed, summarized here and developped in the appendix.

Conjecture 20. *There is no algorithm that can distinguish - with high probability and in polynomial time - a graph with independent edges with probability $1/2$, from a graph in which a subset of k vertices out of n is connected with probability $q > 1/2$; if $k \leq n^\beta$ for $\beta < 1/2$.*

Consider the vectors of explanatory variables $X_i = N^{1/4} e_i$, $i = 1, ..., n$ and assume without loss of generality that the observed set of edges Ω in the matrix logistic regression model consists of the interactions of

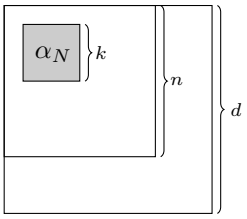
$\mathcal{G}_k^{\alpha_N} \ni \Theta =$ 

Figure 2: The construction of matrices $\mathcal{G}_k^{\alpha_N}$ used in the reduction scheme.

the n nodes X_i, i.e. it holds $N = |\Omega| = \binom{n}{2}$. It follows from the matrix logistic regression modelling assumption (2.1) that the Erdös-Rényi graph $G(n, 1/2)$ corresponds to a random graph associated with the matrix $\Theta_0 = 0 \in \mathbf{R}^{d\times d}$. Let $\mathcal{G}_l(k)$ be a subset of $\mathcal{P}_{k,1}(M)$ with a fixed support l of the block. In addition, let $\mathcal{G}_k^{\alpha_N} \subset \mathcal{P}_{k,1}(M)$ be a subset consisting of the matrices $\Theta_l \in \mathcal{G}_l(k), l = 1, ..., K$, $K = \binom{n}{k}$ such that all elements in the block of a matrix Θ_l equal some $\alpha_N = \alpha/\sqrt{N} > 0$, see Figure 2. Then we have

$$\mathbf{P}\big((i,j) \in E | X_i, X_j\big) = \frac{1}{1 + e^{-X_i^\top \Theta X_j}} = \frac{1}{1+e^{-\alpha}}\,,$$

for all $\Theta \in \mathcal{G}_k^{\alpha_N}$. Therefore, the testing problem

$$H_0 : Y \sim \mathbf{P}_{\Theta_0} \quad \text{vs.} \quad H_1 : Y \sim \mathbf{P}_\Theta, \Theta \in \mathcal{G}_k^{\alpha_N}\,, \quad (4.1)$$

where $Y \in \{0,1\}^N$ is the adjacency vector of binary responses in the matrix logistic regression model, is reduced to the dense subgraph detection problem with $q = 1/(1+e^{-\alpha})$. This reduction scheme suggests that the computational lower bound for separating the hypotheses in the dense subgraph detection problem mimics the computational lower bound for separating the hypotheses in (4.1) in the matrix logistic regression model. The following theorem exploits this fact in order to establish a computational lower bound of order k^2/N for estimating the matrix $\Theta_\star \in \mathcal{P}_{k,r}(M)$.

Theorem 21. *Let $\mathcal{F}_k$ be any class of matrices containing $\mathcal{G}_k^{\alpha_N} \cup \Theta_0$ from the reduction scheme. Let $c > 0$ be a positive constant and $f(k,d,N)$ be a real-valued function satisfying $f(k,d,N) \le ck^2/N$ for $k = k_n < n^\beta$, $0 < \beta < 1/2$ and a sequence $d = d_n$, for all $n > m_0 \in \mathbb{N}$. If Conjecture 20 holds, for some the design $\mathbb{X}$ that fulfils the block isometry property from Definition 3, there is no estimator of $\Theta_\star \in \mathcal{F}_k$, that attains the rate $f(k,d,N)$ for the Frobenius norm risk, and can be evaluated using a (randomised) polynomial-time algorithm, i.e. for any estimator $\hat\Theta$, computable in polynomial time, there exists a sequence $(k,d,N) = (k_n, d_n, N)$, such that*

$$\frac{1}{f(k,d,N)} \sup_{\Theta_\star \in \mathcal{F}_k} \mathbb{E}\big[\|\hat\Theta - \Theta_\star\|_F^2\big] \to \infty\,, \quad (4.2)$$

as $n \to \infty$. Similarly, for any estimator $\hat\Theta$, computable in polynomial time, there exists a sequence $(k,d,N) = (k_n, d_n, N)$, such that

$$\frac{1}{f(k,d,N)} \sup_{\Theta_\star \in \mathcal{F}_k} \frac{1}{N}\mathbb{E}\big[\|\hat\Sigma - \Sigma_\star\|_{F,\Omega}^2\big] \to \infty\,, \quad (4.3)$$

for the prediction error of estimating $\Sigma_\star = \mathbb{X}^\top \Theta_\star \mathbb{X}$.

Remark 22. *Thus the computational lower bound for estimating the matrix $\Theta_\star$ in the matrix logistic regression model is of order k^2/N compared to the minimax rate of estimation of order $kr/N + (k/N)\log(de/k)$ and the rate of estimation $(k^2/N)\log(d)$ for the Lasso estimator $\hat\Theta_{Lasso}$, cf. Figure 1. Hence the computational gap is most noticeable for the matrices of rank 1. Furthermore, as a simple consequence of this result, the corresponding computational lower bound for the prediction risk of estimating $\Sigma_\star = \mathbb{X}^\top\Theta_\star\mathbb{X}$ is k^2/N as well.*

Acknowledgements

During this work, both authors were at the University of Cambridge. NB was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 647812). QB was supported by an Isaac Newton Trust Early Career Support Scheme and by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *arXiv:1703.10146*.

Abbe, E. and Sandon, C. (2015). Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap. *arXiv:1512.09080*.

Abramovich, F. and Grinshtein, V. (2016). Model selection and minimax estimation in generalized linear models. *IEEE Transactions on Information Theory*, 62(6):3721–3730.

Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2011). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88.

Agarwal, A. (2012). *Computational Trade-offs in Statistical Learning*. Theses, University of California, Berkeley.

Alon, N. K. M. and Sudakov, B. (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference "Ran-*

dom Structures and Algorithms" (Poznan, 1997), volume 13, pages 457–466.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414.

Banks, J., Moore, C., Neeman, J., and Netrapalli, P. (2016). Information-theoretic thresholds for community detection in sparse networks. *arXiv:1601.02658*.

Bellet, A., Habrard, A., and Sebban, M. (2014). A survey on metric learning for feature vectors and structured data.

Berthet, Q. and Chandrasekaran, V. (2016). Resource allocation for statistical estimation. *Proceedings of the IEEE*, 104(1):115–125.

Berthet, Q. and Ellenberg, J. (2015). Detection of planted solutions for flat satisfiability problems.

Berthet, Q. and Perchet, V. (2017). Fast rates for bandit optimization with upper-confidence frank-wolfe. *NIPS 2017, to appear*.

Berthet, Q. and Rigollet, P. (2013a). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066.

Berthet, Q. and Rigollet, P. (2013b). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815.

Berthet, Q., Rigollet, P., and Srivastava, P. (2016). Exact recovery in the ising blockmodel.

Biau, G. and Bleakly, K. (2008). Statistical inference on graphs. *Statistics and decisions*, 24(2):209–232.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73.

Bruer, J. J., Tropp, J. A., Cevher, V., and Becker, S. R. (2015). Designing statistical estimators that balance sample size, risk, and computational cost. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):612–624.

Bubeck, S., Ding, J., Eldan, R., and Rácz, M. (2014). Testing for high-dimensional geometry in random graphs.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194.

Bunea, F., She, Y., and Wegkamp, M. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5):2359–2388.

Candes, E. J. and Tao, T. (2005). Decoding by Linear Programming. *IEEE Trans. Information Theory*, 51:4203–4215.

Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*.

Chen, Y. (2015). Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923.

Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A., and Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *J. Mach. Learn. Res.*, 10:747–776.

Chen, Y. and Xu, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research*, 17(27):1–57.

Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106.

Devroye, L., György, A., Lugosi, G., and Udina, F. (2011). High-dimensional random geometric graphs and their clique number. *Electronic Communications in Probability*, 16(90):2481–2508.

Fan, J., Gong, W., and Zhu, Z. (2017). Generalized high-dimensional trace regression via nuclear norm regularization. *arXiv preprint arXiv:1710.08083*.

Fan, J., Wang, W., and Zhu, Z. (2016). Robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*.

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

Gao, C., Ma, Z., and Zhou, H. (2017). Sparse cca: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*.

Giraud, C. (2011). Low rank multivariate regression. *Electronic Journal of Statistics*, 5:775–799.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Hajek, B., Wu, Y., and Xu, J. (2015). Computational lower bounds for community detection on random graphs. In *Proceedings of The 28th Conference on*

Learning Theory, volume 40 of *Proceedings of Machine Learning Research*, pages 899–928.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Jain, L., Jamieson, K., and Nowak, R. (2016). Finite sample prediction and recovery bounds for ordinal embedding.

Klopp, O. and Tsybakov, A., a. V. N. (2015). Oracle inequalities for network models and sparse graphon estimation. *Annals of Statistics (to appear)*. arXiv:1507.04118.

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

Kučera, L. (1995). Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57(2-3):193–212. Combinatorial optimization 1992 (CO92) (Oxford).

Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA. ACM.

Ma, Z. and Wu, Y. (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45.

Massoulié, L. (2014). Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289.

Mossel, E., Neeman, J., and Sly, A. (2013). A proof of the block model threshold conjecture. *arXiv:1311.4115*.

Mossel, E., Neeman, J., and Sly, A. (2015). Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461.

Negahban, S. and Wainwright, M. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

Nickl, R. and van de Geer, S. (2013). Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876.

Oymak, S., Jalali, A., Fazel, M., Eldar, Y., and Hassibi, B. (2015). Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908.

Papa, G., Bellet, A., and Clémençon, S. (2016). On graph reconstruction via empirical risk minimization: Fast learning rates and scalability. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 694–702. Curran Associates, Inc.

Penrose, M. (2003). *Random Geometric Graphs*. Oxford University Press.

Richard, E., Savalle, P., and Vayatis, N. (2012). Estimation of simultaneously sparse and low-rank matrices. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1351–1358.

Rigollet, P. (2012). Kullback-leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2):639–665.

Rohde, A. and Tsybakov, A. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Traonmilin, Y. and Gribonval, R. (2015). Stable recovery of low-dimensional cones in hilbert spaces: One rip to rule them all.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Wang, T., Berthet, Q., and Plan, Y. (2016a). Average-case hardness of rip certification. In *Proceedings of*

the 30th International Conference on Neural Information Processing Systems, NIPS'16, pages 3826–3834, USA. Curran Associates Inc.

Wang, T., Berthet, Q., and Samworth, R. (2016b). Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press.

Wolfe, P. and Olhede, S. (2013). Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*.

Yu, H., Braun, P., and Yildirim, M. (2008). High quality binary protein interaction map of the yeast interactome network. *Science (New York, NY)*, 322(322):104–110.

Zhang, A. and Dong, X. (2017). Tensor svd: Statistical and computational limits. *arXiv preprint arXiv:1703.02724*.

Zhang, Y., Levina, E., and Zhu, J. (2015). Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*.

Zhang, Y., Wainwright, M., and Jordan, M. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. 35.