Optimal Link Prediction with Matrix Logistic Regression

NICOLAI BALDIN[*] AND QUENTIN BERTHET[†]

University of Cambridge

Abstract. We consider the problem of link prediction, 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 analysed in a high-dimensional regime under a structural assumption. The minimax rate for the Frobenius-norm risk is established and a combinatorial estimator based on the penalised maximum likelihood approach is shown to achieve it. Furthermore, it is shown that this rate cannot be attained by any (randomised) algorithm computable in polynomial time under a computational complexity assumption.

Primary 62J02, 62C20; secondary 68Q17.
Key words and phrases: Link prediction, Logistic regression, Computational lower bounds, Planted clique.

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.

The fundamental problem of link prediction is therefore naturally linked with statistical estimation: the objective is to understand, through a generative model, why different vertices are connected or not, and to generalise these observations to the rest of the graph.

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 (Alon and Sudakov, 1998; Kučera, 1995), the stochastic block model (Holland et al., 1983)—see Abbe (2017) for a recent survey of a very active line of work (Abbe and Sandon, 2015; Banks et al., 2016; Decelle et al., 2011; Massoulié,

2014; Mossel et al., 2013, 2015), the Ising blockmodel (Berthet et al., 2016), random geometric graphs – see Penrose (2003) for an introduction and (Bubeck et al., 2014; Devroye et al., 2011) for recent developments in statistics, or metric-based learning (Bellet et al., 2014; Chen et al., 2009) and ordinal embedings (Jain et al., 2016).

In supervised regression problems on the other hand, the focus is on under-standing a fundamental mechanism, formalized as the link between two variables. The objective is to learn how an explanatory variable X allows to predict a re-sponse Y, i.e. to find the unknown function f that best approximates the rela-tionship $Y \approx f(X)$. 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 (Hoff et al., 2002; Holland and Leinhardt, 1981).

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 link prediction, using the principle of regression for 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, Gao et al. (2015); Wolfe and Olhede (2013), for graphon estimation, Biau and Bleakly (2008); Papa et al. (2016) for graph reconstruction and Bickel and Chen (2009) for modularity analysis.

Link prediction can be useful in any application where data can be gathered about the nodes of a network. One of the most obvious motivations is in social networks, in order to model social interactions. With access to side information about each member of a social network, the objective is to understand the mech-anisms of connection between members: shared interests, differences in artistic tastes or political opinion (Wasserman and Faust, 1994). This can also be applied to citation networks, or in the natural sciences to biological networks of interac-tions between molecules or proteins (Madeira and Oliveira, 2004; Yu et al., 2008). 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, e.g. where opposites attract.

The focus on a high-dimensional setting is another aspect of this work that is also motivated by modern applications of statistics: data is often collected without discernment and the ambient dimension d can be much larger than the sample size. This setting is common in regression problems: the underlying model is often actually very simple, to reflect the fact that only a small number of measured parameters are relevant to the problem at hand, and that the intrinsic dimension is much smaller. This is usually handled through an assumption on the rank, sparsity, or regularity of a parameter. Here this needs to be adapted to a model with two covariates (explanatory variables) and a structural assumption is made

in order to reflect this nature of our problem.

We therefore decide to tackle link prediction 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 couples (i, j), denoted by Ω.

To convey the general idea of a simple dependency 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 1. We talk of link prediction as this is the legacy name but we focus almost entirely on the problem of estimating $\Theta_\star$.

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 contribution: This work is organized in the following manner: We give a formal description of the problem in Section 1, as well as a discussion of our assumptions and links with related work. Section 2 collects our main statistical results. We propose an estimator $\hat{\Theta}$ based on the penalised maximum likelihood approach and analyse its performance in Section 2.1 in terms of non-asymptotic rate of estimation. We show that it attains the minimax rate of estimation over simultaneously block-sparse and low-rank matrices $\Theta_\star$, but is not computationally tractable. In Section 2.2, we provide a convex relaxation of the problem which is in essence the *Lasso* estimator applied to a vectorised version of the problem. The link prediction task is covered in Section 2.3. A matching minimax lower bound for the rate of estimation is given in Section 2.4. Furthermore, we show in Section 3 that the minimax rate cannot be attained by a (randomised) polynomial-time algorithm, and we identify a corresponding computational lower bound. The proof of this bound is based on a reduction scheme from the so-called *dense subgraph detection problem*. Technical proofs are deferred to the appendix. Our findings are depicted in Figure 1.

Notation: For any positive integer n, we denote by $[n]$ the set $\{1, \ldots, n\}$ and by $[[n]]$ the set of couples 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
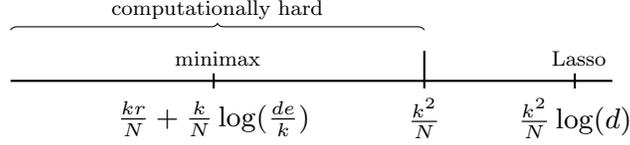
Figure 1: The computational and statistical boundaries for estimation and prediction in the matrix logistic regression model. Here k denotes the sparsity of $\Theta_\star$ and r its rank, while N is the number of observed edges in the network.

$\|B\|_{F,\Omega}$ 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 $g(x)$.

1. PROBLEM DESCRIPTION

1.1 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

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

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* (1.1). *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

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

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*. 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

$$(1.3) \qquad X_i^\top \Theta_\star X_j = \mathbf{Tr}(X_j X_i^\top \Theta_\star) = \langle \mathbf{vec}(X_j X_i^\top), \mathbf{vec}(\Theta_\star) \rangle \,.$$

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 *matrix logistic regression* in order to simplify the notation of the explanatory variables and our model assumptions on $\Theta_\star$, that are specific to matrices.

1.2 Comparaison with 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 (1.3), 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. Abramovich and Grinshtein, 2016; Bach, 2010; Bunea, 2008; Meier et al., 2008; Rigollet, 2012; van de Geer, 2008, 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 Gao et al. (2015); Klopp and Tsybakov (2015); Wolfe and Olhede (2013); 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 (1.1) of the present paper is in fact very close to the trace regression model, as it follows from the representation (1.3). 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 Fan et al. (2016); Koltchinskii et al. (2011); Negahban and Wainwright (2011); Rohde and Tsybakov (2011) 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 $V_1, \ldots, V_n$ 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 (Bellet et al., 2014; Chen et al., 2009) and references therein for a comprehensive survey of metric learning methods.

1.3 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 dimensionality, 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

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

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).

1.4 Explanatory variables

As mentioned above, this problem is different from tasks such as metric learning, where the objective is to estimate the X_i with no side information. Here they 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*

$$N\big(1 - \Delta_{\Omega,s}(\mathbb{X})\big)\|B\|_F^2 \le \|\mathbb{X}^\top B\,\mathbb{X}\|_{F,\Omega}^2 \le N\big(1 + \Delta_{\Omega,s}(\mathbb{X})\big)\|B\|_F^2\,,$$

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

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*

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

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

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

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

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

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 $\|\mathbb{X}^\top B\,\mathbb{X}\|_{F,\Omega}^2 = \|\mathbb{D}_\Omega \boldsymbol{vec}(B)\|_2^2$, and $\|\boldsymbol{vec}(B)\|_2^2 = \|B\|_F^2$.

$\square$

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

$$(1.4) \qquad \mathcal{L}(M) := \sigma'(M) = \sigma(M)\big(1 - \sigma(M)\big),$$

is lower bounded away from zero, and we have

$$(1.5) \qquad \inf_{\Theta \in \mathcal{P}(M)} \sigma'(X_i^\top \Theta X_j) \geq \mathcal{L}(M) > 0,$$

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 Abramovich and Grinshtein (2016); van de Geer (2008). Similarly to Bach (2010), Assumption 6 can be shown to be redundant for minimax optimal prediction, because the log-likelihood function in the matrix logistic regression model satisfies the so-called self-concordant property. Our analysis to follow can be combined with an analysis similar to Bach (2010) to get rid of the assumption for minimax optimal prediction.*

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*

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

1.4.1 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.

2. 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 $\Theta \in \mathbf{S}^d$

$$\ell(\Theta) = \ell(\Theta_\star) - \sum_{(i,j)\in\Omega} \mathsf{KL}(\pi_{ij}(\Theta_\star), \pi_{ij}(\Theta))$$

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

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.

2.1 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

$$(2.1) \qquad \hat{\Theta} \in \operatorname*{argmin}_{\Theta\in\mathcal{P}(M)} \Big\{ -\ell_Y(\Theta) + p(\Theta) \Big\} \,,$$

with a penalty p defined as

$$(2.2) \qquad p(\Theta) = g(\operatorname{rank}(\Theta), \|\Theta\|_{0,0}) \,, \quad \text{and} \quad g(R, K) = cKR + cK \log\Big(\frac{de}{K}\Big) \,,$$

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* (2.2) *with the constants $c \geq c_1/\mathcal{L}$, $c_1 > 1$, $\mathcal{L}$ given in* (1.4). *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

$$(2.3) \qquad \sup_{\Theta_\star \in \mathcal{P}_{k,r}(M)} \frac{1}{N} \mathbb{E}[\mathsf{KL}(\mathbf{P}_{\Theta_\star}, \mathbf{P}_{\hat{\Theta}})] \leq C_1 \frac{kr}{N} + C_1 \frac{k}{N} \log\left(\frac{de}{k}\right),$$

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 1.4.1 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*

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

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*

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

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

REMARK 14. *The penalty (2.2) 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 (2.1) yields a suboptimal estimator as it follows from Theorem 13.*

2.2 Convex relaxation

In practice, computation of the estimator (2.1) 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. However, they are not feasible in a high-dimensional setting either. We here consider the following estimator

$$(2.4) \qquad \hat{\Theta}_{Lasso} = \underset{\Theta \in \mathbf{S}^d}{\operatorname{argmin}}\{-\ell_Y(\Theta) + \lambda\|\Theta\|_{1,1}\},$$

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 15. *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 (2.4) satisfies*

$$(2.5) \qquad \sup_{\Theta_\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\hat{\Theta}_{Lasso} - \Theta_\star\|_F^2\big] \leq \frac{C_5}{\mathcal{L}(M)\big(1 - \Delta_{\Omega,2k}(\mathbb{X})\big)}\frac{k^2}{N}\log d\,,$$

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. Bunea et al. (2012); Giraud (2011); Koltchinskii et al. (2011); 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 3, 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).

2.3 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 16. *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 $\hat{\Sigma} = \mathbb{X}^\top \hat{\Theta} \mathbb{X} \in \mathbf{R}^{n\times n}$:*

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

with the constant C_1 from (2.3). 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 $\Sigma_\star = \mathbb{X}^\top \Theta_\star \mathbb{X} \in \mathbf{R}^{n\times n}$.

2.4 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 17.   *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 18.   *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 $\Sigma_\star = \mathbb{X}^\top \Theta_\star \mathbb{X} \in \mathbf{R}^{n \times n}$.*

3. COMPUTATIONAL LOWER BOUNDS

In this section, we investigate whether the lower bound in Theorem 17 can be achieved with an estimator computable in polynomial time. The fastest rate of estimation attained by a (randomised) polynomial-time algorithm in the worst-case scenario is usually referred to as a *computational lower bound*. Recently, the gap between computational and statistical lower bounds has attracted a lot of attention in the statistical community. We refer to Berthet and Rigollet (2013a,b); Chen (2015); Chen and Xu (2016); Gao et al. (2017); Hajek et al. (2015); Ma and Wu (2015); Wang et al. (2016b); Zhang and Dong (2017) 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.

3.1 The dense subgraph detection problem

Although our work is related to the study of graphs, we recall for absolute clarity the following notions from graph theory. A *graph* $G = (V, E)$ is a non-empty set V of *vertices*, together with a set E of distinct unordered pairs $\{i, j\}$ with $i, j \in V$, $i \neq j$. Each element $\{i, j\}$ of E is an edge and joins i to j. The vertices of an edge are called its endpoints. We consider only undirected graphs with neither loops nor multiple edges. A graph is called *complete* if every pair of distinct vertices is connected. A graph $G' = (V', E')$ is a subgraph of a graph

$G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. A subgraph C is called a *clique* if it is complete. The problem of detecting a maximum clique, or all cliques, the so-called Clique problem, in a given graph is known to be NP-complete, cf. Karp (1972).

The Planted Clique problem, motivated as an average case version of the Clique problem, can be formalised as a decision problem over random graphs, parametrised by the number of vertices n and the size of the subgraph k. Let $\mathbb{G}_n$ denote the collection of all graphs with n vertices and $G(n, 1/2)$ denote distribution of Erdös-Rényi random graphs, uniform on $\mathbb{G}_n$, where each edge is drawn independently at random with probability $1/2$. For any $k \in \{1, ..., n\}$ and $q \in (1/2, 1]$, let $G(n, 1/2, k, q)$ be a distribution on $\mathbb{G}_n$ constructed by first picking k vertices independently at random and connecting all edges in-between with probability q, and then joining each remaining pair of distinct vertices by an edge independently at random with probability $1/2$. Formally, the Planted Clique problem refers to the hypothesis testing problem of

$$(3.1) \qquad H_0 : A \sim G(n, 1/2) \quad \text{vs.} \quad H_1 : A \sim G(n, 1/2, k, 1),$$

based on observing an adjacency matrix $A \in \mathbf{R}^{n \times n}$ of a random graph drawn from either $G(n, 1/2)$ or $G(n, 1/2, k, 1)$.

One of the main properties of the Erdös-Rényi random graph were studied in Erdős and Rényi (1959), as well as in Grimmett and McDiarmid (1975), who in particular proved that the size of the largest clique in $G(n, 1/2)$ is asymptotically close to $2 \log_2 n$ almost surely. On the other hand, Alon et al. (1998) proposed a spectral method that for $k > c\sqrt{n}$ detects a planted clique with high probability in polynomial time. Hence the most intriguing regime for k is

$$(3.2) \qquad 2 \log_2 n \le k \le c\sqrt{n}.$$

The conjecture that no polynomial-time algorithm exists for distinguishing between hypotheses in (3.1) in the regime (3.2) with probability tending to 1 as $n \to \infty$ is the famous Planted Clique conjecture in complexity theory. Its variations have been used extensively as computational hardness assumptions in statistical problems, see Berthet and Rigollet (2013b); Cai and Wu (2018); Gao et al. (2017); Wang et al. (2016b).

The Planted Clique problem can be reduced to the so-called dense subgraph detection problem of testing the null hypothesis in (3.1) against the alternative $H_1 : A \sim G(n, 1/2, k, q)$, where $q \in (1/2, 1]$. This is clearly a computationally harder problem. In this paper, we assume the following variation of the Planted Clique conjecture which is used to establish a computational lower bound in the matrix logistic regression model.

CONJECTURE 19 (The dense subgraph detection conjecture). *For any sequence $k = k_n$ such that $k \le n^\beta$ for some $0 < \beta < 1/2$, and any $q \in (1/2, 1]$, there is no (randomised) polynomial-time algorithm that can correctly identify the dense subgraph with probability tending to 1 as $n \to \infty$, i.e. for any sequence of (randomised) polynomial-time tests $(\psi_n : \mathbb{G}_n \to \{0, 1\})_n$, we have*

$$\liminf_{n \to \infty} \left\{ \mathbf{P}_0(\psi_n(A) = 1) + \mathbf{P}_1(\psi_n(A) = 0) \right\} \ge 1/3.$$

3.2 Reduction to the dense subgraph detection problem and a computational lower bound

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 the n nodes X_i, i.e. it holds $N = |\Omega| = \binom{n}{2}$. It follows from the matrix logistic regression modelling assumption (1.1) 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

$$(3.3) \qquad\qquad H_0 : Y \sim \mathbf{P}_{\Theta_0} \quad \text{vs.} \quad H_1 : Y \sim \mathbf{P}_\Theta, \Theta \in \mathcal{G}_k^{\alpha_N} \,,$$

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 (3.3) 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 20.   *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 $f(k, d, N) \leq ck^2/N$ 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 19 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*

$$(3.4) \qquad\qquad \frac{1}{f(k, d, N)} \sup_{\Theta_\star \in \mathcal{F}_k} \mathbb{E}\big[\|\hat{\Theta} - \Theta_\star\|_F^2\big] \to \infty \,,$$

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

$$(3.5) \qquad\qquad \frac{1}{f(k, d, N)} \sup_{\Theta_\star \in \mathcal{F}_k} \frac{1}{N} \mathbb{E}\big[\|\hat{\Sigma} - \Sigma_\star\|_{F,\Omega}^2\big] \to \infty \,,$$

for the prediction error of estimating $\Sigma_\star = \mathbb{X}^\top \Theta_\star \mathbb{X}$.
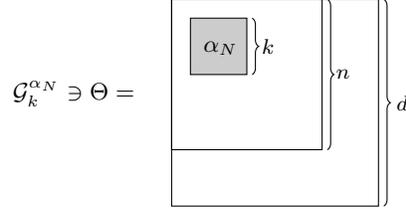
$$\mathcal{G}_k^{\alpha_N} \ni \Theta =$$

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

REMARK 21.   *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.*

PROOF. We here provide a proof of the computational lower bound on the prediction error (3.5) for convenience. The bound on the estimation error (3.4) is straightforward to show by utilizing the block isometry property. Assume that there exists a hypothetical estimator $\hat{\Theta}$ computable in polynomial time that attains the rate $f(k, d, N)$ for the prediction error, i.e. such that it holds that

$$\limsup_{n\to\infty} \frac{1}{f(k,d,N)} \sup_{\Theta_\star \in \mathcal{F}_k} \frac{1}{N} \mathbb{E}\big[\|\mathbb{X}^\top(\hat{\Theta} - \Theta_\star)\mathbb{X}\|_{F,\Omega}^2\big] \le b < \infty,$$

for all sequences $(k, d, N) = (k_n, d_n, N)$ and a constant b. Then by Markov's inequality, we have

$$(3.6) \qquad \frac{1}{N}\|\mathbb{X}^\top(\hat{\Theta} - \Theta_\star)\mathbb{X}\|_{F,\Omega}^2 \le u f(k, d, N),$$

for some numeric constant $u > 0$ with probability $1 - b/u$ for all $\Theta_\star \in \mathcal{F}_k$. Following the reduction scheme, we consider the design vectors $X_i = N^{1/4} e_i$, $i = 1, ..., n$ and the subset of edges Ω, such that

$$(3.7) \qquad \frac{1}{N}\|\mathbb{X}^\top(\hat{\Theta} - \Theta_\star)\mathbb{X}\|_{F,\Omega}^2 = \sum_{(i,j)\in\Omega} (\hat{\Theta}_{ij} - \Theta_{\star ij})^2 = \|\hat{\Theta} - \Theta_\star\|_{F,\Omega}^2,$$

for any $\Theta_\star \in \mathcal{G}_k^{\alpha_N}$. Note, that the design vectors $X_i = N^{1/4} e_i$, $i = 1, ..., n$ clearly satisfy Assumption 6. Thus, in order to separate the hypotheses

$$(3.8) \qquad H_0 : Y \sim \mathbf{P}_0 \quad \text{vs.} \quad H_1 : Y \sim \mathbf{P}_\Theta, \Theta \in \mathcal{G}_k^{\alpha_N},$$

it is natural to employ the following test

$$(3.9) \qquad \psi(Y) = \mathbf{1}\big(\|\hat{\Theta}\|_{F,\Omega} \ge \tau_{d,k}(u)\big),$$

where $\tau_{d,k}^2(u) = u f(k, d, N)$. The type I error of this test is controlled automatically due to (3.6) and (3.7), $\mathbf{P}_0(\psi = 1) \le b/u$. For the type II error, we obtain

$$\sup_{\Theta\in\mathcal{G}_k^{\alpha_N}} \mathbf{P}_\Theta(\psi = 0) = \sup_{\Theta\in\mathcal{G}_k^{\alpha_N}} \mathbf{P}_\Theta\big(\|\hat{\Theta}\|_{F,\Omega} < \tau_{d,k}(u)\big)$$

$$\le \sup_{\Theta\in\mathcal{G}_k^{\alpha_N}} \mathbf{P}_\Theta\big(\|\hat{\Theta} - \Theta\|_{F,\Omega}^2 > \|\Theta\|_{F,\Omega}^2 - \tau_{d,k}^2(u)\big) \le b/u,$$

provided that
$$k(k-1)\alpha_N^2/2 \geq 2\tau_{d,k}^2(u) = 2uf(k,d,N),$$

which is true in the regime $k \leq n^\beta$, $\beta < 1/2$, and $\alpha^2 \geq 4u/c$, (hence $\alpha_N^2 \geq 4u/(cN)$) by the definition of the function $f(k,d,N)$. Putting the pieces together, we obtain

$$\limsup_{n\to\infty} \left\{ \mathbf{P}_0(\psi(Y) = 1) + \sup_{\Theta \in \mathcal{G}_k^{\alpha_N}} \mathbf{P}_\Theta(\psi(Y) = 0) \right\} \leq 2b/u < 1/3,$$

for $u > 6b$. Hence, the test (3.9) separates the hypotheses (3.8). This contradicts Conjecture 19 and implies (3.5).

$\square$

4. CONCLUDING REMARKS

Our results shed further light on the emerging topic of statistical and computational trade-offs in high-dimensional estimation. The matrix logistic regression model is very natural to study the connection between statistical accuracy and computational efficiency as the model is based on the study of a generative model for random graphs. It is also an extension of lower bound *for all* statistical procedures to a model with covariates, the first of its kind.

Our fundings suggest that the block-sparsity is a limiting model selection criterion for polynomial-time estimation in the logistic regression model. That is, imposing further structure, like an additional low-rank constraint, and thus reducing the number of studied models yields an expected gain in the minimax rate, but that gain can never be achieved by a polynomial-time algorithm. In this setting, this implies that with a larger parameter space, while the statistical rates might be worse, they might be closer to those that are computationally achievable. As an illustration, both efficient and minimax optimal estimation is possible for estimating sparse vectors in the high-dimensional linear regression model, see, for example, SLOPE for achieving the exact minimax rate in Bellec et al. (2018); Bogdan et al. (2015), extending upon previous results on the Danzig selector and Lasso in Bickel et al. (2009); Candes and Tao (2007).

The logistic regression is also a representative of a large class of generalised linear models. Furthermore, the proof of the minimax lower bound on the rate of estimation in Theorem 17 can be extended to all generalised linear models. The combinatorial estimator (2.1) can well be used to achieve the minimax rate. The computational lower bound then becomes a delicate issue. A more sophisticated reduction scheme is needed to relate the dense subgraph detection problem to an appropriate testing problem for a generalised linear model. Approaching this question might require notions of noise discretisation and Le Cam equivalence studied in Ma and Wu (2015).

An interesting question is whether it is possible to adopt polynomial-time algorithms available for detecting a dense subgraph for estimating the target matrix in the logistic regression model in all sparsity regimes. A common idea behind those algorithms is to search a dense subgraph over the vertices of a high degree and thus substantially reduce the number of compared models of subgraphs. The network we observe in the logistic regression model is generated by a sparse matrix. We may still observe a fully connected network which is

generated by a small block in the target matrix. Therefore, it is not yet clear how to adopt algorithms for dense subgraph detection to submatrix detection. It remains an open question to establish whether these results can be extended to any design matrix, and all parameter regimes.

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** 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** 61–88.

AGARWAL, A. (2012). *Computational Trade-offs in Statistical Learning*. Theses, University of California, Berkeley. http://research.microsoft.com/en-us/um/people/alekha/thesismain.pdf.

ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. *Random Structures and Algorithms* **13** 457–466.

ALON, N. K. M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*, vol. 13.

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* .

BELLEC, P., LECUÉ, G. and TSYBAKOV, A. (2018). Slope meets lasso: improved oracle bounds and optimality. *Annals of Statistics (to appear)* ArXiv:1605.08651.

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** 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*.

BERTHET, Q. and RIGOLLET, P. (2013b). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics* **41** 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** 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** 21068–21073.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37** 1705–1732.
URL https://doi.org/10.1214/08-AOS620

BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields* **138** 33–73.

BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. (2015). Slope—adaptive variable selection via convex optimization. *The Annals of Applied Atatistics* **9** 1103.

BOUSQUET, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique* **334** 495–500.

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** 612–624.

BUBECK, S., DING, J., ELDAN, R. and RÁCZ, M. (2014). Testing for high-dimensional geometry in random graphs .
URL https://arxiv.org/abs/1411.5713

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** 2359–2388.

CAI, T. and WU, Y. (2018). Statistical and computational limits for sparse matrix detection. *arXiv preprint arXiv:1801.00518* .

CANDES, E. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory* **57** 2342–2359.

CANDES, E. and TAO, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351.
URL https://doi.org/10.1214/009053606000001523

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** 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** 1–57.
URL http://jmlr.org/papers/v17/14-330.html

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** 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** 2481–2508.

DUDLEY, R. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis* **1** 290 – 330.

ERDŐS, P. and RÉNYI, A. (1959). On random graphs. *Publicationes Mathematicae* **6** 290–297.

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** 2624–2652.

GAO, C., MA, Z. and ZHOU, H. (2017). Sparse cca: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565* .

GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

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.

GRIMMETT, G. and MCDIARMID, C. (1975). On colouring random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 77. Cambridge Univ Press.

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*, vol. 40 of *Proceedings of Machine Learning Research*.

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** 1090–1098.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 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** 33–50.

JAIN, L., JAMIESON, K. and NOWAK, R. (2016). Finite sample prediction and recovery bounds for ordinal embedding .

KARP, R. (1972). *Reducibility among combinatorial problems*. Springer.

KLOPP, O. and TSYBAKOV, A. V. N., A. (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** 2302–2329.

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

MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics* **43** 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** 24–45.

MASSART, P. (2007). *Concentration inequalities and model selection*, vol. 6. Springer.

MASSOULIÉ, L. (2014). Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. 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** 53–71.

MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation* **28** 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** 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** 1069–1097.

NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 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** 2886–2908.

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

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)*.

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

ROHDE, A. and TSYBAKOV, A. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* **39** 887–930.

TRAONMILIN, Y. and GRIBONVAL, R. (2015). Stable recovery of low-dimensional cones in hilbert spaces: One rip to rule them all .
URL https://arxiv.org/abs/1510.00504

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation.* 1st ed. Springer.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36** 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, Curran Associates Inc., USA. URL http://dl.acm.org/citation.cfm?id=3157382.3157525

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

APPENDIX A: PROOFS

A.1 Some geometric properties of the likelihood

Let us recall the *stochastic component* of the likelihood function

$$\zeta(\Theta) = \ell_Y(\Theta) - \ell(\Theta) = \sum_{(i,j)\in\Omega} \left(Y_{(i,j)} - \pi_{ij}(\Theta_\star)\right) X_i^\top \Theta X_j\,,$$

which is a linear function in Θ. The deviation of the gradient $\nabla\zeta$ of the stochastic component is governed by the deviation of the independent Bernoulli random variables $\varepsilon_{i,j} = Y_{(i,j)} - \mathbb{E}[Y_{(i,j)}] = Y_{(i,j)} - \pi_{ij}(\Theta_\star)$, $(i,j) \in \Omega$. Let us introduce an upper triangular matrix $\mathcal{E}_\Omega = (\varepsilon_{i,j})_{(i,j)\in\Omega}$ with zeros on the complement set Ω^c. In this notation, we have $\zeta(\Theta) = \langle\!\langle \zeta, \Theta \rangle\!\rangle_F$, with

$$\nabla\zeta = \sum_{(i,j)\in\Omega} \varepsilon_{i,j} X_j X_i^\top = \mathbb{X}\,\mathcal{E}_\Omega\,\mathbb{X}^\top \in \mathbf{R}^{d\times d}\,.$$

In particular, $\nabla\zeta$ is sub-Gaussian with parameter $\sum_{(i,j)\in\Omega} \|X_j X_i^\top\|_F^2/4 = \|\mathbb{X}^\top\mathbb{X}\|_{F,\Omega}^2/4$, i.e. it holds for the moment generating function of $\langle\!\langle \zeta, B \rangle\!\rangle_F$ for any $B \in \mathbf{R}^{d\times d}$ and $\sigma^2 = 1/4$,

$$\varphi_{\langle\!\langle \zeta, B\rangle\!\rangle_F}(t) := \mathbb{E}\left[\exp(t\langle\!\langle \zeta, B \rangle\!\rangle_F)\right] = \prod_{(i,j)\in\Omega} \mathbb{E}\left[\exp(t\varepsilon_{i,j}\langle\!\langle X_j X_i^\top, B \rangle\!\rangle_F)\right]$$

$$(A.1) \qquad \leq \prod_{(i,j)\in\Omega} \exp\left(t^2\sigma^2 \langle\!\langle X_j X_i^\top, B \rangle\!\rangle_F^2/2\right) = \exp\left(t\sigma^2\|\mathbb{X}^\top B\mathbb{X}\|_{F,\Omega}^2/2\right).$$

We shall be frequently using versions of the following inequality, which is based on the fact that $\nabla\ell(\Theta_\star) = 0 \in \mathbf{R}^{d\times d}$, the Taylor expansion and (1.5), and holds for any $\Theta \in \mathcal{P}(M)$,

$$\ell(\Theta_\star) - \ell(\Theta) = \frac{1}{2} \sum_{(i,j)\in\Omega} \left(\sigma'(X_i^\top \Theta_0 X_j)\langle\!\langle X_j X_i^\top, \Theta_\star - \Theta \rangle\!\rangle^2\right)$$

$$(A.2) \qquad \geq \frac{\mathcal{L}}{2} \sum_{(i,j)\in\Omega} \langle\!\langle X_j X_i^\top, \Theta_\star - \Theta \rangle\!\rangle_F^2 = \frac{\mathcal{L}}{2}\|\mathbb{X}^\top(\Theta_\star - \Theta)\mathbb{X}\|_{F,\Omega}^2\,,$$

where $\Theta_0 \in [\Theta, \Theta_\star]$ element-wise. Furthermore, using that $\sup_{t\in\mathbf{R}} \sigma'(t) \leq 1/4$, we obtain for all $\Theta \in \mathcal{P}(M)$

$$\ell(\Theta_\star) - \ell(\Theta) \leq \frac{1}{8}\|\mathbb{X}^\top(\Theta_\star - \Theta)\mathbb{X}\|_{F,\Omega}^2\,.$$

We shall also be using the bounds

$$(A.3) \qquad \max_{(i,j)\in\Omega} \left(\varepsilon_{i,j} X_i^T(\Theta - \Theta_\star)X_j\right) \leq \|\mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\|_{F,\Omega}\,, \quad \text{a.s.}\,,$$

$$(A.4) \qquad \mathrm{Var}\left(\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle_F\right) \leq \frac{1}{4}\|\mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\|_{F,\Omega}^2\,.$$

A.2 Entropy bounds for some classes of matrices

Recall that an ε-*net* of a bounded subset $\mathbf{K}$ of some metric space with a metric ρ is a collection $\{K_1, ..., K_{N_\varepsilon}\} \in \mathbf{K}$ such that for each $K \in \mathbf{K}$, there exists $i \in \{1, ..., N_\varepsilon\}$ such that $\rho(K, K_i) \le \varepsilon$. The ε-*covering number* $N(\varepsilon, \mathbf{K}, \rho)$ is the cardinality of the smallest ε-net. The ε-*entropy* of the class $\mathbf{K}$ is defined by $H(\varepsilon, \mathbf{K}, \rho) = \log_2 N(\varepsilon, \mathbf{K}, \rho)$. The following statement is adapted from Lemma 3.1 in Candes and Plan (2011).

LEMMA 22. *Let* $\mathbb{T}_0 := \{\Theta \in \mathbf{R}^{k \times k} : \mathbf{rank}(\Theta) \le r, \|\Theta\|_F \le 1\}$. *Then it holds for any* $\varepsilon > 0$

$$H(\varepsilon, \mathbb{T}_0, \| \cdot \|_F) \le \big((2k+1)r+1\big) \log\big(\frac{9}{\varepsilon}\big).$$

A.3 Proof of Theorem 10 and Theorem 16

It suffices to show the following uniform deviation inequality

$$(\mathrm{A.5}) \qquad \sup_{\Theta_\star \in \mathcal{P}_{k,r}(M)} \mathbf{P}_{\Theta_\star}\big(\ell(\Theta_\star) - \ell(\hat{\Theta}) + p(\hat{\Theta}) > 2p(\Theta_\star) + R_t^2\big) \le e^{-cR_t},$$

for any $R_t > 0$ and some numeric constant $c > 0$. Indeed, then taking $R_t^2 = p(\Theta_\star)$, it follows that $\ell(\Theta_\star) - \ell(\hat{\Theta}) \le 3p(\Theta_\star)$ uniformly for all $\Theta_\star$ in the considered class with probability at least $1 - e^{-c\sqrt{p(\Theta_\star)}}$. The upper bound (2.3) of Theorem 10 the follows directly integrating the deviation inequality (A.5), while the upper bound on the prediction error in Theorem 16 further follows using (A.2) and the smoothness of the logistic function, $\sup_{t \in \mathbf{R}} \sigma'(t) \le 1/4$. Define

$$(\mathrm{A.6}) \qquad \tau^2(\Theta; \Theta_\star) := \ell(\Theta_\star) - \ell(\Theta) + p(\Theta), \ G_R(\Theta_\star) := \{\Theta : \tau(\Theta; \Theta_\star) \le R\}.$$

The inequality (A.5) clearly holds on the event $\{\tau^2(\hat{\Theta}; \Theta_\star) \le 2p(\Theta_\star)\}$. In view of $\ell_Y(\hat{\Theta}) - p(\hat{\Theta}) \ge \ell_Y(\Theta_\star) - p(\Theta_\star)$, we have on the complement:

$$\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\hat{\Theta} - \Theta_\star)\mathbb{X} \rangle\!\rangle \ge \ell(\Theta_\star) - \ell(\hat{\Theta}) + p(\hat{\Theta}) - p(\Theta_\star) \ge \frac{1}{2}\tau^2(\hat{\Theta}; \Theta_\star).$$

Therefore, for any $\Theta_\star \in \mathcal{P}_{k,r}(M)$, we have

$$\mathbf{P}_{\Theta_\star}\big(\tau^2(\hat{\Theta}; \Theta_\star) > 2p(\Theta_\star) + R_t^2\big) \le \mathbf{P}_{\Theta_\star}\Big(\sup_{\tau(\Theta; \Theta_\star) \ge R_t} \frac{\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle}{\tau^2(\Theta; \Theta_\star)} \ge \frac{1}{2}\Big).$$

We now apply the so-called "peeling device" (or "slicing" as it sometimes called in the literature). The idea is to "slice" the set $\tau(\Theta; \Theta_\star) \ge R_t$ into pieces on which the penalty term $p(\Theta)$ is fixed and the term $\ell(\Theta_\star) - \ell(\Theta)$ is bounded. It follows,

$$\mathbf{P}_{\Theta_\star}\Big(\sup_{\tau(\Theta; \Theta_\star) \ge R_t} \frac{\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle}{\tau^2(\Theta; \Theta_\star)} \ge \frac{1}{2}\Big)$$

$$(\mathrm{A.7})$$

$$\le \sum_{K=1}^d \sum_{R=1}^K \sum_{s=1}^\infty \mathbf{P}_{\Theta_\star}\Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta_\star) \\ k(\Theta)=K, \mathbf{rank}(\Theta)=R}} \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle \ge \frac{1}{8}2^{2s} R_t^2\Big).$$

On the set $\{\Theta \in G_{2^s R_t}(\Theta_\star), k(\Theta) = K, \mathbf{rank}(\Theta) = R\}$, it holds by the definitions (A.6)

$$\ell(\Theta_\star) - \ell(\Theta) \le 2^{2s} R_t^2 - p(K, R)\,,$$

and therefore using (A.2), this implies

(A.8) $\quad \|\mathbb{X}^\top (\Theta_\star - \Theta)\mathbb{X}\|_{F,\Omega} \le Z(K, R, s)\,, \quad Z^2(K, R, s) = \frac{2}{\mathcal{L}} \big(2^{2s} R_t^2 - p(K, R) \big)\,.$

Let us fix the location of the block, that is the support of a matrix $\Theta' \in \mathcal{G}_1 := \{\Theta \in \mathbf{R}^{d \times d} : k(\Theta) = K, \mathbf{rank}(\Theta) = R\}$ belongs to the upper-left block of size $K \times K$. Then following the lines of the proof of Lemma 22 and using the singular value decomposition, we derive

$$H(\varepsilon, \{\mathbb{X}^\top \Theta' \mathbb{X} : \Theta' \in \mathcal{G}_1, \|\mathbb{X}^\top \Theta' \mathbb{X}\|_{F,\Omega} \le B\}, \| \cdot \|_{F,\Omega}) \le \big((2K+1)R+1 \big) \log \big(\frac{9B}{\varepsilon} \big)\,.$$

Consequently, for the set $\mathbb{T} := \{\mathbb{X}^\top (\Theta - \Theta_\star)\mathbb{X} : \Theta \in \mathbf{R}^{d \times d}, \mathbf{rank}(\Theta) = R, k(\Theta) = K, \|\mathbb{X}^\top (\Theta_\star - \Theta)\mathbb{X}\|_{F,\Omega} \le Z(K, R, s)\}$, we obtain

$$H(\varepsilon, \mathbb{T}, \| \cdot \|_{F,\Omega}) \le \big((2K+1)R+1 \big) \log \big(\frac{9Z(K, R, s)}{\varepsilon} \big) + K \log \big(\frac{de}{K} \big)\,.$$

Denote $t(K, R) := \sqrt{KR} + \sqrt{K \log \big(\frac{de}{K} \big)}$. By Dudley's entropy integral bound, see Dudley (1967) and Giné and Nickl (2016) for a more recent reference, we then have

$$\mathbb{E} \big[\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta_\star) \\ k(\Theta)=K, \mathbf{rank}(\Theta)=R}} \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top (\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle \big] \le C' \int_0^{Z(K,R,s)} \sqrt{H(\varepsilon, \mathbb{T}, \| \cdot \|_\Omega)} d\varepsilon$$

$$\le C'' \sqrt{kr} \int_0^{9Z(K,R,s)} \sqrt{\log \big(\frac{9Z(K, R, s)}{\varepsilon} \big)} d\varepsilon + 9C'' Z(K, R, s) \sqrt{K \log \big(\frac{de}{K} \big)}$$

$$\le C Z(K, R, s) t(K, R)\,,$$

for some universal constant $C > 0$. Furthermore, by Bousquet's version of Talagrand's inequality, see Theorem A.10, in view of the bounds (A.3) and (A.4), we have for all $u > 0$

$$\mathbf{P}_{\Theta_\star} \Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta_\star) \\ k(\Theta)=K, \mathbf{rank}(\Theta)=R}} \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top (\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle \ge C Z(K, R, s) t(K, R)$$

$$+ \sqrt{\Big(\frac{1}{2} Z^2(K, R, s) + 4C Z^2(K, R, s) t(K, R) \Big) u} + \frac{Z(K, R, s)u}{3} \Big) \le e^{-u}\,.$$

Taking $u(K, R, s) := \mathcal{L}^{1/2} Z(K, R, s) + \mathcal{L}^{-1/2} t(K, R) + 2 \log d$ and using inequalities $\sqrt{c_1 + c_2} \le \sqrt{c_1} + \sqrt{c_2}$ and $\sqrt{c_1 c_2} \le \frac{1}{2} (c_1 \varepsilon + \frac{c_2}{\varepsilon})$, which hold for any $c_1, c_2, \varepsilon > 0$, we obtain

$$\mathbf{P}_{\Theta_\star} \Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta_\star) \\ k(\Theta)=K, \mathbf{rank}(\Theta)=R}} \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top (\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle$$

$$\ge \frac{1}{16} \mathcal{L} Z^2(K, R, s) + C_1^2 t^2(K, R)/\mathcal{L} \Big) \le e^{-u(K,R,s)}\,,$$

for some numeric constant $C_1 > 0$. Plugging this back into (A.7) and using (A.8), we obtain

$$\mathbf{P}_{\Theta_\star}\Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta_\star) \\ k(\Theta)=K, \mathbf{rank}(\Theta)=R}} \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\rangle\!\rangle \geq \frac{1}{8}2^{2s}R_t^2\Big) \leq e^{-u(K,R,s)}\,,$$

for some numeric constant $C_2 > 0$, provided that

$$\text{(A.9)} \quad \frac{1}{16}\mathcal{L}Z^2(K,R,s) + \frac{C_1^2}{\mathcal{L}}t^2(K,R) \leq \frac{1}{8}2^{2s}R_t^2 = \frac{1}{16}\mathcal{L}Z^2(K,R,s) + 8p(K,R)\,,$$

which is satisfied for $p(K,R) \geq (C_1^2/\mathcal{L})t^2(K,R)$. Therefore,

$$\mathbf{P}_{\Theta_\star}\Big(\sup_{\tau(\Theta;\Theta_\star)\geq R_t} \frac{\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\rangle\!\rangle}{\tau^2(\Theta;\Theta_\star)} \geq \frac{1}{2}\Big) \leq \sum_{K=1}^d \sum_{R=1}^K \sum_{s=1}^\infty e^{-u(K,R,s)} \leq e^{-cR_t}\,,$$

for some numeric constants $c > 0$ using (A.9), which concludes the proof.

The following prominent result is due to Bousquet (2002).

THEOREM 23 (Bousquet's version of Talagrand's inequality). *Let $(B, \mathcal{B})$ be a measurable space and let $\varepsilon_1, ..., \varepsilon_n$ be independent B-valued random variables. Let $\mathcal{F}$ be a countable set of measurable real-valued functions on B such that $f(\varepsilon_i) \leq b < \infty$ a.s. and $\mathbb{E}f(\varepsilon_i) = 0$ for all $i = 1, ..., n$, $f \in \mathcal{F}$. Let*

$$S := \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(\varepsilon_i)\,, \qquad v := \sup_{f\in\mathcal{F}}\sum_{i=1}^n \mathbb{E}[f^2(\varepsilon_i)].$$

Then for all $u > 0$, it holds that

$$\text{(A.10)} \qquad \mathbf{P}\Big(S - \mathbb{E}[S] \geq \sqrt{2(v + 2b\mathbb{E}[S])u} + \frac{bu}{3}\Big) \leq e^{-u}\,.$$

A.4 Proof of Theorem 13

For the MLE $\hat{\Theta}_{k,r}$, it clearly holds $\ell_Y(\hat{\Theta}_{k,r}) \geq \ell_Y(\Theta_\star)$ implying

$$\ell(\Theta_\star) - \ell(\hat{\Theta}_{k,r}) \leq \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\hat{\Theta}_{k,r} - \Theta_\star)\mathbb{X}\rangle\!\rangle.$$

Furthermore, in view of (A.2), we derive

$$\text{(A.11)} \qquad \frac{\mathcal{L}}{2}\|\mathbb{X}^\top(\hat{\Theta}_{k,r} - \Theta_\star)\mathbb{X}\|_{F,\Omega} \leq \frac{\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\hat{\Theta}_{k,r} - \Theta_\star)\mathbb{X}\rangle\!\rangle}{\|\mathbb{X}^\top(\hat{\Theta}_{k,r} - \Theta_\star)\mathbb{X}\|_{F,\Omega}}$$

$$\text{(A.12)} \qquad\qquad\qquad \leq \sup_{\Theta\in\mathcal{P}_{k,r}(M)} \frac{\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\rangle\!\rangle}{\|\mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\|_{F,\Omega}}\,.$$

Following the lines of Section A.3, by Dudley's integral we next obtain

$$\mathbb{E}\Big(\sup_{\Theta\in\mathcal{P}_{k,r}(M)} \frac{\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\rangle\!\rangle}{\|\mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\|_{F,\Omega}}\Big) \leq c\sqrt{kr} + c\sqrt{k\log\Big(\frac{de}{k}\Big)}\,,$$

for some universal constant $c > 0$. Plugging this bound back into (A.12) and using the block isometry property yields the desired assertion.

A.5 Proof of Theorem 17

PROOF. The proof is split into two parts. First, we show a lower bound of the order kr and then a lower bound of the order $k \log(de/k)$. A simple inequality $(a+b)/2 \le \max\{a,b\}$ for all $a, b > 0$ then completes the proof. Both parts of the proof exploit a version of remarkable Fano's inequality given in Proposition 24 to follow, cf. Section 2.7.1 in Tsybakov (2008).

1. A bound kr. The proof of this bound is similar to the proof of a minimax lower bound for estimating a low-rank matrix in the trace-norm regression model given in Theorem 5 in Koltchinskii et al. (2011). For the sake of completeness, we provide the details here. Consider a subclass of matrices

$$\mathcal{C} = \big\{ \ A \in \mathbf{R}^{k \times r} : a_{i,j} = \{0, \alpha_N\}, 1 \le i \le k, 1 \le j \le r \ \big\} \,,$$
$$\alpha_N^2 = \frac{\gamma \log 2}{1 + \Delta_{\Omega, 2k}(\mathbb{X})} \frac{r}{2kN} \,,$$

where $\gamma > 0$ is a positive constant, $\Delta_{\Omega, 2k}(\mathbb{X}) > 0$ is the the block isometry constant from Definition 3 and $\lfloor x \rfloor$ denotes the integer part of x. Further define

$$\mathcal{B}(\mathcal{C}) = \big\{ \frac{1}{2}(A + A^\top) : A = (\tilde{A}| \cdots |\tilde{A}|O) \in \mathbf{R}^{k \times k}, \tilde{A} \in \mathcal{C} \big\} \,,$$

where O denotes the $k \times (k - r\lfloor k/r \rfloor)$ zero matrix. By construction, any matrix $\Theta \in \mathcal{B}(\mathcal{C})$ is symmetric, has rank at most r with entries bounded by α_N. Applying a standard version of the Varshamov-Gilbert lemma, see Lemma 2.9 in Tsybakov (2008), there exists a subset $\mathcal{B}^\circ \subset \mathcal{B}(\mathcal{C})$ of cardinality $\mathsf{card}(\mathcal{B}^\circ) \ge 2^{kr/16} + 1$ such that

$$\frac{kr}{16}\big(\frac{\alpha_N}{2}\big)^2 \big\lfloor \frac{k}{r} \big\rfloor \le \|\Theta_u - \Theta_v\|_F^2 \le k^2 \alpha_N^2 \,,$$

for all $\Theta_u, \Theta_v \in \mathcal{B}^\circ$. Thus $\mathcal{B}^\circ$ is a 2δ-separated set in the Frobenius metric with $\delta^2 = \frac{kr}{64}\big(\frac{\alpha_N}{2}\big)^2 \lfloor \frac{k}{r} \rfloor$. The Kullback-Leibler divergence between the measures $\mathbf{P}_{\Theta_u}$ and $\mathbf{P}_{\Theta_v}$, $\Theta_u, \Theta_v \in \mathcal{B}^\circ$, $u \neq v$, is upper bounded as

$$\mathsf{KL}(\mathbf{P}_{\Theta_u}, \mathbf{P}_{\Theta_v}) = \mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_u)] - \mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_v)] \le \frac{1}{8} \sum_{(i,j) \in \Omega} \langle\!\langle X_j X_i^\top, \Theta_u - \Theta_v \rangle\!\rangle_F^2$$

$$\le \frac{1 + \Delta_{\Omega, 2k}(\mathbb{X})}{8} k^2 \alpha_N^2 N \,.$$

Taking $\gamma > 0$ small enough, we obtain

$$\frac{1 + \Delta_{\Omega, 2k}(\mathbb{X})}{8} k^2 \alpha_N^2 N + \log 2 = \frac{kr}{16} \gamma \log 2 + \log 2 = \log(2^{\frac{kr}{16}\gamma + 1}) < \log(2^{kr/16} + 1) \,,$$

which, in view of Proposition 24, yields the desired lower bound.

2. A bound $k \log(de/k)$. Let $K = \binom{d}{k}$ and consider the set $\mathcal{G}_k^{\alpha_N} \subset \mathcal{P}_{k,1}(M)$ from the reduction scheme in Section 3.2 with

$$\alpha_N^2 = \frac{4\gamma \log 2}{kN(1 + \Delta_{\Omega, 2k}(\mathbb{X}))} \log\big(\frac{de}{k}\big) \,,$$

where $\gamma > 0$ is a positive constant. Using simple calculations, we then have $(2k-1)\alpha_N^2 \le \|\Theta_u - \Theta_v\|_F^2 \le 2k^2\alpha_N^2$ for all $\Theta_u, \Theta_v \in \mathcal{G}_k^{\alpha_N}$, $u \neq v$. Furthermore, according to Lemma 25 to follow, there exists a subset $\mathcal{G}_k^{\alpha_N, 0} \subset \mathcal{G}_k^{\alpha_N}$ such that

$$c_0 k^2 \alpha_N^2 \le \|\Theta_u - \Theta_v\|_F^2 \le 2k^2 \alpha_N^2 \,,$$

and of cardinality $\mathsf{card}(\mathcal{G}_k^{\alpha_N,0}) \geq 2^{\rho k \log(de/k)} + 1$ for some $\rho > 0$ depending on a constant $c_0 > 0$ and independent of k and d. Thus $\mathcal{G}_k^{\alpha_N,0}$ is a 2δ-separated set in the Frobenius metric with $\delta^2 = c_0 k^2 \alpha_N^2/4$. The Kullback-Leibler divergence between the measures $\mathbf{P}_{\Theta_u}$ and $\mathbf{P}_{\Theta_v}$, $\Theta_u, \Theta_v \in \mathcal{G}_k^{\alpha_N,0}$, $u \neq v$, is upper bounded as

$$\mathsf{KL}(\mathbf{P}_{\Theta_u}, \mathbf{P}_{\Theta_v}) = \mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_u)] - \mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_v)] \leq \frac{1}{8} \sum_{(i,j) \in \Omega} \langle\!\langle X_j X_i^\top, \Theta_u - \Theta_v \rangle\!\rangle_F^2$$

$$\leq \frac{1 + \Delta_{\Omega,2k}(\mathbb{X})}{4} k^2 \alpha_N^2 N \,,$$

for all $u \neq v$ and $\Delta_{\Omega,2k}(\mathbb{X}) > 0$ from Definition 3. As in the first part of the proof, taking $\gamma > 0$ small enough, we obtain

$$\frac{1 + \Delta_{\Omega,2k}(\mathbb{X})}{4} k^2 \alpha_N^2 N + \log 2 = k\gamma \log(2) \log\left(\frac{de}{k}\right) + \log 2 = \log(2^{k\gamma \log(de/k)+1})$$

$$< \log(2^{\rho k \log(de/k)} + 1) \,.$$

The desired lower bound then follows from Proposition 24. $\qquad\square$

PROPOSITION 24 (Fano's method). *Let $\{\Theta_1, ..., \Theta_J\}$ be a 2δ-separated set in $\mathbf{R}^{d \times d}$ in the Frobenius metric, meaning that $\|\Theta_k - \Theta_l\|_F \geq 2\delta$ for all elements Θ_k, Θ_l, $l \neq k$ in the set. Then for any increasing and measurable function $F : [0,\infty) \to [0,\infty)$, the minimax risk is lower bounded as*

$$\inf_{\hat{\Theta}} \sup_{\Theta} \mathbb{E}_{\mathbf{P}_\Theta}\big[F(\|\hat{\Theta} - \Theta\|_F)\big] \geq F(\delta)\Big(1 - \frac{\sum_{u,v}\mathsf{KL}(\mathbf{P}_{\Theta_u}, \mathbf{P}_{\Theta_v})/J^2 + \log 2}{\log J}\Big) \,.$$

LEMMA 25 (Variant of the Varshamov-Gilbert lemma). *Let $\mathcal{G} \subset \mathcal{P}_{k,1}(M)$ be a set of $\{0,1\}^{d \times d}$ symmetric block-sparse matrices with the size of the block k, where $k \leq \alpha\beta d$ for some $\alpha, \beta \in (0,1)$. Denote $K = \binom{d}{k}$ the cardinality of $\mathcal{G}$ and $\rho_H(E, E') = \sum_{i,j} \mathbf{1}(E_{i,j} \neq E'_{i,j})$ the Hamming distance between two matrices $E, E' \in \mathcal{G}$. Then there exists a subset $\mathcal{G}^0 = \{E^{(0)}, ..., E^{(J)}\} \subset \mathcal{G}$ of cardinality*

$$\log J := \log(\mathsf{card}(\mathcal{G}^0)) \geq \rho k \log\left(\frac{de}{k}\right) \,,$$

where $\rho = \frac{\alpha}{-\log(\alpha\beta)}(-\log\beta + \beta - 1)$ such that

$$\rho_H(E^{(k)}, E^{(l)}) \geq ck^2 \,,$$

for all $k \neq l$ where $c = 2(1 - \alpha^2) \in (0,2)$.

PROOF. Let $E^{(0)} = \{0\}^{k \times k}$, $D = ck^2$, and construct the set $\mathcal{E}_1 = \{E \in \mathcal{G} : \rho_H(E^{(0)}, E) > D\}$. Next, pick any $E^{(1)} \in \mathcal{E}_1$ and proceed iteratively so that for a matrix $E^{(j)} \in \mathcal{E}_j$ we construct the set

$$\mathcal{E}_{j+1} = \{E \in \mathcal{E}_j : \rho(E^{(j)}, E) > D\} \,.$$

Let J denote the last index j for which $\mathcal{E}_j \neq \varnothing$. It remains to bound the cardinality J of the constructed set $\mathcal{G}^0 = \{E^{(0)}, ..., E^{(J)}\}$. For this, we consider the cardinality n_j of the subset $\{\mathcal{E}_j \setminus \mathcal{E}_{j+1}\}$:

$$n_j := \#\{\mathcal{E}_j \setminus \mathcal{E}_{j+1}\} \leq \#\{E \in \mathcal{G} : \rho_H(E^{(j)}, E) \leq D\} \,.$$

For all $E, E' \in \mathcal{G}$, we have

$$\rho_H(E, E') = 2(k^2 - (k - m)^2),$$

where $m \in [0, k]$ corresponds to the number of distinct columns of E (or E'). Solving the quadratic equation (A.5) for $\rho_H(E, E') = D = ck^2$ we obtain

$$m_D = k(1 - \sqrt{1 - c/2}),$$

for the maximum number of distinct columns of a block-sparse matrix E (and E') such that $\rho_H(E, E') \leq D = ck^2$ for $c \in [0, 2]$. For instance, in order to get the distance between matrices $2k^2$, i.e. $c = 2$ we need to shift all the k columns (and consequently rows) and so the number of distinct columns of a matrix is $m = k$, and in order to get the minimal possible distance $4k - 2$, i.e. $c = (4k - 2)/k^2$ we need to shift only one column and a corresponding row, i.e. $m = 1$. Therefore, for n_j in (A.5), we have

$$n_j \leq \#\{E \in \mathcal{G} : \rho_H(E^{(j)}, E) \leq D\} = \sum_{i=0}^{m_D} \binom{k}{i}\binom{d-k}{i} = \sum_{i=k-m_D}^{k} \binom{k}{i}\binom{d-k}{k-i}.$$

Together with an evident equality $\sum_{j=0}^{J} n_j = K = \binom{d}{k}$, this implies

$$\sum_{i=k-m_D}^{k} \binom{k}{i}\binom{d-k}{k-i} \bigg/ \binom{d}{k} \geq \frac{1}{J+1}.$$

Note that taking $m_D = k$, which as we have seen corresponds to $c = 2$, we have a trivial bound $J + 1 \geq 1$ using Vandermonde's convolution. Furthermore, the expression on the left-hand side in (A.5) is exactly the probability $\mathbf{P}(X \geq k - m_D) = \mathbf{P}(X \geq k\alpha)$ for $\alpha = \sqrt{1 - c/2}$, where the variable X follows the hypergeometric distribution $H(d, k, k/d)$. The rest of the proof is based on applying Chernoff's inequality and follows the scheme of the proof of Lemma 4.10 in Massart (2007). $\qquad\square$

NICOLAI BALDIN
DEPARTMENT OF PURE MATHEMATICS
AND MATHEMATICAL STATISTICS,
UNIVERSITY OF CAMBRIDGE
WILBEFORCE ROAD
CAMBRIDGE, CB3 0WB, UK
(n.baldin@statslab.cam.ac.uk)

QUENTIN BERTHET
DEPARTMENT OF PURE MATHEMATICS
AND MATHEMATICAL STATISTICS,
UNIVERSITY OF CAMBRIDGE
WILBEFORCE ROAD
CAMBRIDGE, CB3 0WB, UK
(q.berthet@statslab.cam.ac.uk)