Adaptive confidence sets for matrix completion

Alexandra Carpentier, Olga Klopp, Matthias Löffler, Richard Nickl

Adaptive confidence sets for matrix completion

February 3, 2017

Alexandra Carpentier, *Universität Potsdam*[1]

Olga Klopp, *University Paris Ouest*[2]

Matthias Löffler and Richard Nickl, *University of Cambridge*[3]

Abstract

In the present paper we study the problem of existence of honest and adaptive confidence sets for matrix completion. We consider two statistical models: the trace regression model and the Bernoulli model. In the trace regression model, we show that honest confidence sets that adapt to the unknown rank of the matrix exist even when the error variance is unknown. Contrary to this, we prove that in the Bernoulli model, honest and adaptive confidence sets exist only when the error variance is known a priori. In the course of our proofs we obtain bounds for the minimax rates of certain composite hypothesis testing problems arising in low rank inference.

Keywords. Low rank recovery, confidence sets, adaptivity, matrix completion, unknown variance, minimax hypothesis testing.

1 Introduction

In matrix completion we observe n noisy entries of a data matrix $M = (M_{ij}) \in \mathbb{R}^{m_1 \times m_2}$, and we aim at doing inference on M. In a typical situation of interest, n is much smaller than $m_1 m_2$, the total number of entries. This problem arises in many applications such as recommender systems and collaborative filtering [3, 21], genomics [18] or sensor localization [37]. Two statistical models have been proposed in the matrix completion literature: the *trace-regression model* (e.g. [9, 27, 29, 31, 36]) and the *'Bernoulli model'* (e.g. [10, 17, 28]).

In the *trace-regression model* we observe n pairs (X_i, Y_i^{tr}) satisfying

$$Y_i^{tr} = \langle X_i, M \rangle + \epsilon_i = \mathrm{tr}(X_i^T M) + \epsilon_i, \quad i = 1, \dots, n, \tag{1.1}$$

where (ϵ_i) is a noise vector. The random matrices $X_i \in \mathbb{R}^{m_1 \times m_2}$ are independent of the ϵ_i's, chosen uniformly at random from the set

$$\mathcal{B} = \left\{ e_j(m_1) e_k^T(m_2),\ 1 \le j \le m_1,\ 1 \le k \le m_2 \right\}, \tag{1.2}$$

were the $e_j(s)$ are the canonical basis vectors of $\mathbb{R}^s$. In this model Y_i^{tr} returns the noisy value of the entry of M corresponding to the random position X_i.

In the *Bernoulli model* each entry of $M + E$, where $E = (\epsilon_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ is a matrix of random errors, is observed independently of the other entries with probability $p = n/(m_1 m_2)$. More precisely, if $n \le m_1 m_2$ is given and B_{ij} are i.i.d. Bernoulli random variables of parameter p independent of the ϵ_{ij}'s, we observe

$$Y_{ij}^{Ber} = B_{ij} \left(M_{ij} + \epsilon_{ij} \right), \quad 1 \le i \le m_1, 1 \le j \le m_2. \tag{1.3}$$

[1]Institut für Mathematik, carpentier@maths.uni-potsdam.de
[2]MODALX, kloppolga@math.cnrs.fr
[3]Statistical Laboratory, Centre for Mathematical Sciences, m.loffler@statslab.cam.ac.uk, r.nickl@statslab.cam.ac.uk

The major difference between these models is that in the trace-regression model multiple sampling of a particular entry is possible whereas in the Bernoulli model each entry can be sampled at most once. A further difference is that in the trace regression model the number of observations, n, is fixed whereas in the Bernoulli model the number of observations $\hat{n} := \sum_{ij} B_{ij}$ is random with expectation $E\hat{n} = n$. Despite these differences, the results on minimax optimal recovery using computationally efficient algorithms for these two models in the literature are very similar and from a 'parameter estimation' point of view the models appear to be effectively equivalent (see, e.g., [9, 11, 12, 13, 17, 22, 26, 27, 29, 31, 34]). *A key insight of the present paper is that for the construction of optimal confidence sets, these models are in fact fundamentally different, at least when the noise variance σ^2 is unknown.*

When investigating questions that go beyond mere 'adaptive estimation' of a high-dimensional parameter, such as about the existence of adaptive confidence sets, one can expect to encounter surprising phenomena – and various recent results (see e.g. [2, 6, 7, 19, 23, 25, 30, 32, 33, 35, 38] and Chapter 8.3 in [20]) show that the answers depend on a rather subtle interaction of certain 'information geometric' properties of the model – the material relevant for the present paper is reviewed in Section 2. Many of these results reveal limitations by showing that confidence regions that adapt to the whole parameter space do not exist unless one makes specific 'signal strength' assumptions. For example, Low [30] and Giné and Nickl [19] investigated this question in nonparametric density estimation and Nickl and van de Geer [33] in the sparse high-dimensional regression model.

Next to the challenge of adaptation, the construction of confidence sets in the matrix completion setting is difficult mainly due to two reasons. Firstly, the Restricted Isometry Property (RIP) does not hold, requiring a more involved analysis than in a standard trace regression setting such as in [15]. Moreover, in most practical applications of matrix completion such as movie recommender systems [3, 21] the variance of the errors is not known. Typical constructions of confidence sets in high-dimensions such as χ^2-confidence sets (e.g. [33, 15]) require explicit knowledge of the variance and are thus not feasible. Particularly in the 'Bernoulli model', the problem of unknown variance can be expected to be potentially severe: for the related standard normal means model (without low rank structure and without missing observations) Baraud [2] has shown that in the unknown variance case honest confidence sets of shrinking diameter do not exist, even if the true model is low dimensional. Similarly, in high-dimensional regression Cai and Guo [8] prove the impossibility of constructing adaptive confidence sets for the l_q-loss, $1 \leq q \leq 2$, of adaptive estimators if the variance is unknown.

Our main contributions are as follows: in the trace regression model, even if only an upper bound for the variance of the noise is known, it is shown that practical honest confidence sets exist that have Frobenius-norm diameter that adapts to the unknown rank of M. Contrary to this we prove that such confidence regions cannot exist in the Bernoulli model when the noise variance is unknown, and to complement our findings we also prove that in the Bernoulli model with *known* variance, adaptive confidence regions *do* exist. So while recovery algorithms for matrix completion are not sensitive to the choice of model, the task of uncertainty quantification for these algorithms *is*, and crucially depends on the statistician's ability to estimate the noise variance. For the Bernoulli 'normal means' model our results imply that the lack of availability of 'repeated samples' induces an information-theoretic 'barrier' for inference even in the presence of low rank structure.

This paper is organized as follows: in Subsection 1.1 we formulate the assumptions and collect notation which we use throughout the paper. Then, in Section 2, we review and present general results about the existence of honest and adaptive confidence sets in terms of some information-theoretic quantities that determine the complexity of the adaptation problem at hand. Afterwards we review the literature on minimax estimation in matrix completion problems. In Section 4 we give an explicit construction of honest and adaptive confidence sets in the trace-regression case, adapting a U-statistic approach inspired by Robins and van der Vaart [35] (see also [20], Section 6.4, and [15]). Finally, we present our results for the Bernoulli model in Section 5. First, we derive an upper bound for the minimax rate of testing a low rank hypothesis and deduce from it the existence of honest and adaptive confidence regions in the known variance case. We then derive a lower bound for this testing rate in the unknown variance case, from which we can deduce that honest and adaptive confidence sets over the whole parameter space cannot exist in general. Sections 7-8 contain the proofs of our results.

1.1 Notation & assumptions

By construction, in the Bernoulli model (1.3) the expected number of observations, n, is smaller than the total number of matrix entries, i.e. $n \leq m_1 m_2$. To provide a meaningful comparison we will assume throughout that $n \leq m_1 m_2$ also holds in the trace regression model (1.1). In many applications of matrix completion, such as recommender systems (e.g. [3, 21]) or sensor localization (e.g. [4, 37]) the noise is bounded but not necessarily identically distributed. This is the assumption which we adopt in the present paper. More precisely, we assume that the ϵ_ι are independent random variables that are homoscedastic, have zero mean and are bounded:

Assumption 1.1. *In the models (1.1) and (1.3) with index $\iota = i$ and $\iota = (i,j)$, respectively, we assume $\mathbb{E}(\epsilon_\iota) = 0$, $\mathbb{E}(\epsilon_\iota^2) = \sigma^2$, $\epsilon_\iota \perp\!\!\!\perp \epsilon_\eta$ for $\iota \neq \eta$ and that there exists a positive constant $U > 0$ such that almost surely*

$$\max_\iota |\epsilon_\iota| \leq U.$$

We denote by $M = (M_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ the unknown matrix of interest and define

$$m = \min(m_1, m_2),$$
$$d = m_1 + m_2.$$

For any $l \in \mathbb{N}$ we set $[l] = \{1, \dots, l\}$. Let A, B be matrices in $\mathbb{R}^{m_1 \times m_2}$. We define the matrix scalar product as $\langle A, B \rangle := \mathrm{tr}(A^T B)$. The trace norm of the matrix A is defined as $\|A\|_* := \sum \sigma_j(A)$, the operator norm as $\|A\| := \sigma_1(A)$ and the Frobenius norm as $\|A\|_F^2 := \sum_i \sigma_i^2 = \sum_{i,j} A_{ij}^2$ where $(\sigma_j(A))$ are the singular values of A arranged in decreasing order. Finally $\|A\|_\infty = \max_{i,j} |A_{ij}|$ denotes the largest absolute value of any entry of A. Given a semi-metric $\mathcal{D}$ we define the diameter of a set S by

$$|S|_{\mathcal{D}} := \sup\{\mathcal{D}(x,y) : \quad x, y \in S\}.$$

Furthermore, for $k \in \mathbb{N}_0$ we define the parameter space of rank k matrices with entries bounded by a in absolute value as

$$\mathcal{A}(a,k) := \{A \in \mathbb{R}^{m_1 \times m_2} : \|A\|_\infty \leq a \quad \text{and} \quad \mathrm{rank}(A) \leq k\}. \tag{1.4}$$

Finally, for a subset $\Sigma \subset (0, U]$ we define

$$\mathcal{A}(a,k) \otimes \Sigma := \{(A, \sigma) : \ A \in \mathcal{A}(a,k), \ \sigma \in \Sigma\}.$$

As usual, for sequences a_n and b_n we say $a_n \lesssim b_n$ if there exists a constant C independent of n such that $a_n \leq C \cdot b_n$ for all n. We write $\mathbb{P}_{M,\sigma}$ (and $\mathbb{E}_{M,\sigma}$ for the corresponding expectation) for the distribution of the observations in the models (1.1) or (1.3), respectively.

2 Minimax theory for adaptive confidence sets

In this section we present results about existence of honest and adaptive confidence sets in a general minimax framework. To this end, let $Y = Y^n \sim \mathbb{P}_f^n$ on some measure space $(\Omega_n, \mathcal{B}), n \in \mathbb{N}$, where f is contained in some parameter space $\mathcal{A}$, endowed with a semi-metric $\mathcal{D}$. Let r_n denote the minimax rate of estimation over $\mathcal{A}$, i.e.

$$\inf_{\tilde{f}_n : \Omega_n \to \mathcal{A}} \sup_{f \in \mathcal{A}} \mathbb{E}_f \mathcal{D}(\tilde{f}, f) \asymp r_n(\mathcal{A}).$$

We consider an 'adaptation hypothesis' $\mathcal{A}_0 \subset \mathcal{A}$ characterised by the fact that the minimax rate of estimation in $\mathcal{A}_0$ is of asymptotically smaller order than in $\mathcal{A}$: $r_n(\mathcal{A}_0) = o(r_n(\mathcal{A}))$ as $n \to \infty$. In our matrix inference setting we will choose for $\mathcal{D}$ the distance induced by $\|\cdot\|_F$, for $\mathcal{A}_0, \mathcal{A}$ the parameter spaces $\mathcal{A}(a, k_0) \otimes \Sigma$, $\mathcal{A}(a,k) \otimes \Sigma$ from above, $k_0 = o(k)$ as $\min(n,m) \to \infty$, and data (Y_i, X_i) or (Y_{ij}, B_{ij}) arising from equation (1.1) or (1.3), respectively.

Definition 2.1 (Honest and adaptive confidence sets). *Let $\alpha, \alpha' > 0$ be given. A set $C_n = C_n(Y, \alpha) \subset \mathcal{A}$ is a honest confidence set at level α for the model $\mathcal{A}$ if*

$$\liminf_n \inf_{f \in \mathcal{A}} \mathbb{P}_f^n(f \in C_n) \geq 1 - \alpha. \tag{2.1}$$

Furthermore, we say that C_n is adaptive for the sub-model $\mathcal{A}_0$ at level α' if there exists a constant $K = K(\alpha, \alpha') > 0$ such that

$$\sup_{f \in \mathcal{A}_0} \mathbb{P}_f^n \left(|C_n|_{\mathcal{D}} > K r_n(\mathcal{A}_0) \right) \leq \alpha' \tag{2.2}$$

while still retaining

$$\sup_{f \in \mathcal{A}} \mathbb{P}_f^n \left(|C_n|_{\mathcal{D}} > K r_n(\mathcal{A}) \right) \leq \alpha'. \tag{2.3}$$

We next introduce certain composite testing problems.

Definition 2.2 (Minimax rate of testing & uniformly consistent tests). *Consider the testing problem*

$$H_0 : f \in \mathcal{A}_0 \quad \textit{against} \quad H_1 : f \in \mathcal{A},\ \mathcal{D}(f, \mathcal{A}_0) \geq \rho_n \tag{2.4}$$

where $(\rho_n : n \in \mathbb{N})$ is a sequence of non-negative numbers. We say that ρ_n is the minimax rate of testing for (2.4) if

(i) $\forall \beta > 0$ $\exists$ a constant $L = L(\beta) > 0$ and a test $\Psi_n = \Psi_n(\beta)$, $\Psi_n : \Omega_n \to \{0, 1\}$ such that

$$\sup_{f \in \mathcal{A}_0} \mathbb{E}_f[\Psi_n] + \sup_{f \in \mathcal{A},\ \mathcal{D}(f, \mathcal{A}_0) \geq L\rho_n} \mathbb{E}_f\left[1 - \Psi_n\right] \leq \beta. \tag{2.5}$$

We say that such a test Ψ_n is β-uniformly consistent.

(ii) For some $\beta_0 > 0$ and any sequence $\rho_n^* = o(\rho_n)$ we have

$$\liminf_{n \to \infty} \inf_{\Psi_n : \Omega_n \to \{0,1\}} \left[\sup_{f \in \mathcal{A}_0} \mathbb{E}_f[\Psi_n] + \sup_{f \in \mathcal{A},\ \mathcal{D}(f, \mathcal{A}_0) \geq \rho_n^*} \mathbb{E}_f\left[1 - \Psi_n\right] \right] \geq \beta_0 > 0. \tag{2.6}$$

Theorem 2.1. *Let ρ_n be the minimax rate of testing for the testing problem (2.4) and suppose that $\beta_0 > 0$ is as in (2.6). Suppose that*

$$r_n(\mathcal{A}_0) = o(\rho_n).$$

Then a honest and adaptive confidence set C_n that satisfies (2.1)-(2.3) for any $\alpha, \alpha' > 0$ such that $0 < 2\alpha + \alpha' < \beta_0$ does not exist. In fact if $3\alpha < \beta_0$, then for any honest confidence set C_n that satisfies (2.1) we have that

$$\sup_{f \in \mathcal{A}_0} \mathbb{E}_f |C_n|_{\mathcal{D}} \geq c \rho_n. \tag{2.7}$$

for a constant $c = c(\alpha) > 0$.

The first claim of this theorem is Proposition 8.3.6 in [20]. The lower bound (2.7) also follows from that proof, arguing as in the proof of Theorem 4 in [16].

A converse of Theorem 2.1 also exists, as can be extracted from Proposition 8.3.7 in [20] and an observation in Carpentier (see [14], proof of Theorem 3.5 in Section 6). For this we need the notion of an *oracle*-estimator.

Definition 2.3 (Oracle estimator). *Let $\beta > 0$ be given. We say that an estimator $\hat{f}$ satisfies an oracle inequality at level β if there exists a constant C such that for all $f \in \mathcal{A}$ we have with $\mathbb{P}_f^n$-probability at least $1 - \beta$,*

$$\mathcal{D}(\hat{f}, f) \leq C \inf_{\tilde{\mathcal{A}} \in \{\mathcal{A}, \mathcal{A}_0\}} \left(\mathcal{D}(f, \tilde{\mathcal{A}}) + r_n(\tilde{\mathcal{A}}) \right). \tag{2.8}$$

This is a typical property of adaptive estimators, and is for example in the trace-regression setting fulfilled by the soft-thresholding estimator proposed by Koltchinskii et.al. [29]. The following theorem proves that if the minimax rate of testing is no larger than the minimax rate of estimation in the adaptation hypothesis, then honest adaptive confidence sets do exist. The proof is constructive and yields a confidence set of non-asymptotic coverage at least $1-\alpha$.

Theorem 2.2. *Let $\alpha, \alpha' > 0$ be given. Let ρ_n be the minimax rate of testing for the problem (2.4) such that a $\min(\alpha/2, \alpha')$-uniformly consistent test exists. Assume that $\rho_n \leq C' r_n(\mathcal{A}_0)$ for some constant $C' = C(\alpha, \alpha') > 0$. Moreover, assume that an oracle estimator $\hat{f}$ at level $\alpha/2$ fulfilling (2.8) exists. Then there exists a confidence set C_n that adapts to the sub-model $\mathcal{A}_0$ at level α' satisfying (2.2), (2.3) and that is honest at level α, i.e.,*

$$\sup_{f \in \mathcal{A}} \mathbb{P}_f^n \left(f \notin C_n \right) \leq \alpha.$$

3 Minimax matrix completion

Noisy matrix completion has been extensively studied in several papers starting from Candes and Plan [12], see e.g. [13, 26, 29, 31, 27, 17, 9, 28, 34]). Optimal rates have been achieved under various sets of assumptions. For instance the construction of the estimator (and the resulting upper bound) in [31] requires knowledge of the 'spikiness' ratio of the unknown matrix and leads to sub-optimal rates in the case of sparse matrices. The bounds due to Keshavan et. al. [26] are also only optimal for certain classes of matrices, namely almost square matrices that have a condition number bounded by a constant and fulfil the incoherence condition introduced by [10]. Optimal convergence rates for the classes $\mathcal{A}(a,k)$ of matrices under consideration in the present paper have been obtained by Koltchinskii. et. al. [29] and Klopp [27] for the trace-regression model and by Klopp [28] for the Bernoulli model. For example, in the trace-regression setting, Klopp [27] shows that a constrained Matrix Lasso estimator $\hat{M} := \hat{M}(a, \sigma)$ satisfies with $\mathbb{P}_{M_0,\sigma}$-probability at least $1 - 2/d$

$$\frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} \leq C \frac{kd \log(d)}{n} \quad \text{and} \quad \|M_0 - \hat{M}\|_\infty \leq 2a \tag{3.1}$$

as long as $m \log(d) \leq n \leq d^2 \log(d)$ and where $C = C(\sigma, a) > 0$. Similarly, in the Bernoulli model with noise bounded by U it has been shown in Klopp [28] that an iterative soft thresholding estimator $\hat{M} := \hat{M}(a, \sigma)$ satisfies with $\mathbb{P}_{M_0,\sigma}$-probability at least $1 - 8/d$

$$\frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} \leq C \frac{kd}{n} \quad \text{and} \quad \|M_0 - \hat{M}\|_\infty \leq 2a \tag{3.2}$$

for $n \geq m \log(d)$ and for a constant $C = C(\sigma, a, U) > 0$. Matching lower bounds have also been shown by Koltchinskii. et a. [29] and Klopp [28]. In the trace-regression model with Gaussian noise we have for constants $\beta \in (0,1)$ and $c = c(\sigma, a) > 0$ that

$$\inf_{\hat{M}} \sup_{M_0 \in \mathcal{A}(a,k)} \mathbb{P}_{M_0,\sigma} \left(\frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} > c \frac{kd}{n} \right) \geq \beta.$$

A similar lower bound can be obtained in the Bernoulli setting (see Klopp [28]). These lower and upper bounds imply that for the Frobenius loss and the parameter space $\mathcal{A}(a,k)$ the minimax rate $r_{n,m}(\mathcal{A}(a,k))$ is (at most up to a log-factor) of order

$$\sqrt{m_1 m_2 kd/n}. \tag{3.3}$$

4 Trace Regression Model

We first consider the trace regression model. For the sake of precision we sometimes write M_0 for the 'true parameter' M that has generated the equation (1.1).
For notational simplicity we assume that n is even. Then we can split our observations in two independent sub-samples of equal size $n/2$. In what follows all probabilistic statements are under the distribution $\mathbb{P}$ (with corresponding expectation written $\mathbb{E}$) of the first sub-sample $(Y_i^{tr}, X_i)_{i \leq n/2}$ of size $n/2 \in \mathbb{N}$, conditional on the second sub-sample $(Y_i^{tr}, X_i)_{i > n/2}$, i.e. we have $\mathbb{P}(.) = \mathbb{P}_{M_0,\sigma}(\cdot \,|(Y_i^{tr}, X_i)_{i > n/2})$.

4.1 A non-asymptotic confidence set in the trace regression model with known variance of the errors.

In this case we can adapt the construction of [15]: we first unbiasedly estimate the risk $\|\hat{M} - M_0\|_F^2/(m_1 m_2)$ of a minimax optimal estimator $\hat{M}$ computed from an independent sample (e.g., via sample splitting) by a natural χ^2-statistic (see (4.1)). The construction of an unbiased estimate requires knowledge of σ^2, but when available this estimate, enlarged by natural quantile constants, serves as a good proxy for the diameter of the confidence set C_n centred at $\hat{M}$.

More precisely, using only the second sub-sample $(Y_i^{tr}, X_i)_{i>n/2}$ we compute the matrix lasso estimator from Klopp [27] which achieves the bound (3.1) with probability at least $1 - 2/d$. Then, we freeze $\hat{M}$ and the second sub-sample. We define the following residual sum of squares statistic:

$$\hat{R}_n = \frac{2}{n} \sum_{i \le n/2} (Y_i^{tr} - \langle X_i, \hat{M} \rangle)^2 - \sigma^2. \tag{4.1}$$

Given $\alpha > 0$, let $\xi_{\alpha,\sigma,U} = \sqrt{2}\sigma U \log(\alpha)$, $z_\alpha = \log(3/\alpha)$ and, for a $z > 0$, a fixed constant to be chosen, define the confidence set

$$C_n = \left\{ A \in \mathbb{R}^{m_1 \times m_2} : \frac{\|A - \hat{M}\|_F^2}{m_1 m_2} \le 2 \left(\hat{R}_n + z\frac{d}{n} + \frac{\bar{z} + \xi_{\alpha,\sigma,U}}{\sqrt{n}} \right) \right\}, \tag{4.2}$$

where

$$\bar{z}^2 = \bar{z}^2(\alpha, d, n, \sigma, z) = z_\alpha \sigma^2 \max\left(\frac{3\|A - \hat{M}\|_F^2}{m_1 m_2}, 4zd/n \right).$$

It is not difficult to see (using that $x^2 \lesssim y + x/\sqrt{n}$ implies $x^2 \lesssim y + 1/n$) that

$$\mathbb{E}_{M_0,\sigma} \left[\frac{|C_n|_F^2}{m_1 m_2} \middle| \hat{M} \right] \lesssim \frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} + \frac{zd + \sigma^2 z_{\alpha/3}}{n} + \frac{\xi_{\alpha,\sigma,U}}{\sqrt{n}}. \tag{4.3}$$

Markov's inequality, (4.3) and that $\hat{M}$ is minimax optimal (up to a log-factor) with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1 - 2/d$ as long as $m \log(d) \le n \le d^2 \log(d)$ imply that C_n has an adaptive and up to a log-factor minimax optimal squared diameter with probability $1 - \alpha'$ for any $\alpha' > 2/d$. The following theorem shows that C_n is also a honest confidence set:

Theorem 4.1. *Let $\alpha > 0$, $\alpha' > 2/d$ and suppose that $m \log(d) \le n \le d^2 \log(d)$, that Assumption 1.1 is satisfied and that $\sigma > 0$ is known. Let $C_n = C_n(Y, \alpha, \sigma)$ be given by (4.2) with $z > 0$. Then, for every $n \in \mathbb{N}$ and every $M_0 \in \mathcal{A}(a, m)$,*

$$\mathbb{P}_{M_0,\sigma}(M_0 \in C_n) \ge 1 - \frac{2\alpha}{3} - 2e^{-zd/(11a^2)}.$$

Hence, for any $1 \le k_0 < k \le m$, C_n is a honest and (up to a log-factor) adaptive confidence set at the level α for the model $\mathcal{A}(a, k) \otimes \{\sigma\}$ and adapts to the sub-model $\mathcal{A}(a, k_0) \otimes \{\sigma\}$ at level α'.

The proof of Theorem 4.1 follows the lines of the proof of Theorem 2 in [15] and we omit it here as the unknown variance results considered in the next section straightforwardly imply the known variance results.

4.2 A non-asymptotic confidence set in the trace regression model with unknown error variance.

In this subsection we assume, that the precise knowledge of the noise variance σ is *not* available, although the quantities a, U are available to the statistician (i.e. upper bounds on the matrix entries and on the noise). More precisely we assume that σ belongs to a known set $\Sigma \subset (0, U]$. In applications of matrix completion this is usually a realistic assumption since the entries of M_0 are bounded: For example in a movie recommender system (e.g. [3, 21]) the entries of the observations Y and consequently M_0 and ϵ_i are bounded from above by the best possible rating and below from the worst possible rating.

As the variance is now assumed to be unknown the construction from (4.2) is not feasible anymore since we can not compute the test statistic (4.1). Instead we use a U-statistic approach: As in the previous section, we use the second half of the sample, $(Y_i^{tr}, X_i)_{n/2<i\leq n}$, for constructing a minimax optimal estimator $\hat{M}$ of M that fulfills $\|\hat{M}\|_\infty \leq a$. We use again the matrix lasso estimator from Klopp [27] (with σ replaced by its upper bound U) which achieves (3.1) with probability at least $1-2/d$. In order to construct the confidence set, we will be interested in all pairs of observations (Y_l^{tr}, X_l) and (Y_s^{tr}, X_s) in the first sub-sample with $1 \leq l < s \leq n/2$ such that $X_l = X_s$ (that is, independent measurements of the same matrix entry). For each $(i,j) \in [m_1] \times [m_2]$, let $\mathcal{S}_{(i,j)} = \{k \in \{1,\dots,n/2\} : X_k = e_i(m_1)e_j^T(m_2)\} =: \{a_1 < ... < a_{p_{(i,j)}}\}$ where $p_{(i,j)}$ is the number of times that we observe the entry (i,j). For all indices (i,j) such that $S_{(i,j)} \neq \emptyset$, we form the $\lfloor p_{(i,j)}/2 \rfloor$ couples $(X_{a_1}, X_{a_2}), (X_{a_3}, X_{a_4}), \dots$ etc. We denote by $\mathcal{N}$ the set of all these pairs and let $|\mathcal{N}| = N$ be their number. Re-ordering, we can write $(\tilde{X}_k, Z_k, Z'_k)_{k\leq N}$ where $\tilde{X}_k = X_l = X_s$ for some couple $(X_l, X_s) \in \mathcal{N}$ and $Z_k = Y_l^{tr}$ and $Z'_k = Y_s^{tr}$. That is, using two different samples of the same entry $\tilde{X}_k = X_l = X_s$ we form the observation triples $(\tilde{X}_k, Z_k, Z'_k)$. We use $(\tilde{X}_k, Z_k, Z'_k)_{k\leq N}$ to construct a U-Statistic to estimate the squared Frobenius loss. Contrary to the construction in (4.1) this does not require knowledge of the variance of the errors. We define:

$$\hat{R}_N := \frac{1}{N}\sum_{k=1}^{N}(Z_k - \langle \hat{M}, \tilde{X}_k\rangle)(Z'_k - \langle \hat{M}, \tilde{X}_k\rangle), \tag{4.4}$$

and we set $\hat{R}_N = 0$ if $N = 0$. Note that

$$\mathbb{E}_{M_0,\sigma}\left[\hat{R}_N \middle| \hat{M}, N \geq 1\right] = \frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2}. \tag{4.5}$$

We define the confidence set

$$C_n := \left\{ A \in \mathcal{A}(a,m) : \quad \frac{\|A - \hat{M}\|_F^2}{m_1 m_2} \leq \hat{R}_N + z_{\alpha,N} \right\} \tag{4.6}$$

where the random quantile constant $z_{\alpha,N}$ is defined as

$$z_{\alpha,N} := \frac{U^2 + 4a^2}{\sqrt{N\alpha}} \quad \text{if} \quad N \neq 0 \quad \text{and} \quad z_{\alpha,N} = 4a^2 \quad \text{if} \quad N = 0.$$

The quantity N is random but we can bound it from below with high probability by $n^2/(64m_1m_2)$ as proven in the following lemma.

Lemma 4.1. *For $n \leq m_1 m_2$ we have with probability at least $1 - \exp\left(-n^2/(372m_1m_2)\right)$ that:*

$$N \geq \frac{n^2}{64 m_1 m_2}.$$

Markov's inequality, (4.5), Lemma 4.1 and that $\hat{M}$ achieves the nearly optimal rate (3.1) with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1-2/d$ imply for any $k \leq m$, any $M_0 \in \mathcal{A}(a,k)$, any $\sigma \leq U$, any $\alpha' > 2/d + \exp(-n^2/(372m_1m_2))$ and a large enough constant $C = C(\alpha, \alpha', \sigma, a, U) > 0$ that

$$\mathbb{P}_{M_0,\sigma}\left(\frac{|C_n|_F^2}{m_1 m_2} > C\frac{kd\log(d)}{n}\right) \leq \alpha'. \tag{4.7}$$

Since k is arbitrary this implies that C_n is a confidence set whose $\|\cdot\|_F^2$-diameter adapts to the unknown rank of M_0 without requiring the knowledge of $\sigma \in \Sigma$. The following theorem implies that C_n is also a honest confidence set. Note that our result is non-asymptotic and holds for any triple $(n, m_1, m_2) \in \mathbb{N}^3$ as long as $m \log d \leq n \leq m_1 m_2$.

Theorem 4.2. *Let $\alpha > 0$ be given, assume $m\log(d) \leq n \leq m_1 m_2$ and that Assumption 1.1 is fulfilled. Let $C_n = C_n(Y, \alpha)$ as in (4.6). Then C_n satisfies for any $M_0 \in \mathcal{A}(a,m)$ and any $\sigma \in \Sigma$*

$$\mathbb{P}_{M_0,\sigma}\left(M_0 \in C_n\right) \geq 1 - \alpha.$$

Hence, for any $\alpha' > 2/d + \exp(-n^2/(372m_1m_2))$ and any $1 \leq k_0 < k \leq m$, C_n is a honest confidence set at level α for the model $\mathcal{A}(a,k)\otimes\Sigma$ that adapts (up to a log-factor) to the rank k_0 of any sub-model $\mathcal{A}(a,k_0)\otimes\Sigma$ at level α'.

5 Bernoulli Model

In this section we consider the Bernoulli model (1.3). As before we let $\mathbb{P}_{M,\sigma}$ (and $\mathbb{E}_{M,\sigma}$ for the corresponding expectation) denote the distribution of the data when the parameters are M and σ, and we sometimes write M_0 for the 'true' parameter M for the sake of precision.

5.1 A non-asymptotic confidence set in the Bernoulli model with known variance of the errors.

Here we assume again that $\sigma > 0$ is known. In case of the Bernoulli model we are not able to obtain two independent samples and cannot use the risk estimation approaches from the trace-regression setting. Instead we use the duality between testing and honest and adaptive confidence sets laid out in Section 2. We first determine an upper bound for the minimax rate $\rho = \rho_{n,m}$ of testing the low rank hypothesis

$$H_0 : M \in \mathcal{A}(a, k_0) \quad \text{against} \quad H_1 : M \in \mathcal{A}(a, k), \quad \|M - \mathcal{A}(a, k_0)\|_F^2 \geq \rho^2, \tag{5.1}$$

and then apply Theorem 2.2. As test statistic, we propose an infimum-test which has previously been used by Bull and Nickl [6] and Nickl and van de Geer [33] in density estimation and high-dimensional regression, respectively (see also Section 6.2.4. in [20]). Since $\sigma^2 = \mathbb{E}\epsilon_{ij}^2$ is known we can define the statistic

$$T_n := \inf_{A \in \mathcal{A}(a,k_0)} \left| \frac{1}{\sqrt{2n}} \sum_{i,j} B_{ij} \left((Y_{ij} - A_{ij})^2 - \sigma^2 \right) \right| = \inf_{A \in \mathcal{A}(a,k_0)} \left| \frac{1}{\sqrt{2n}} \sum_{i,j} \left((Y_{ij} - B_{ij}A_{ij})^2 - B_{ij}\sigma^2 \right) \right| \tag{5.2}$$

and choose the quantile constant u_α such that

$$\mathbb{P}_\sigma \left(\frac{1}{\sqrt{2n}} \left| \sum_{i,j} B_{ij}(\epsilon_{ij}^2 - \mathbb{E}\epsilon_{ij}^2) \right| > u_\alpha \right) \leq \alpha/3. \tag{5.3}$$

For example, using Markov's inequality, we obtain

$$\mathbb{P}_\sigma \left(\frac{1}{\sqrt{2n}} \left| \sum_{i,j} B_{ij}(\epsilon_{ij}^2 - \sigma^2) \right| > u_\alpha \right) \leq \frac{1}{2nu_\alpha^2} \sum_{i,j} \mathrm{Var}_\sigma \left(B_{ij}(\epsilon_{ij}^2 - \sigma^2) \right) \leq \frac{\sigma^2(U^2 - \sigma^2)}{2u_\alpha^2}$$

so $u_\alpha = \sigma\sqrt{\left(3(U^2 - \sigma^2)\right)/(2\alpha)}$ is an admissible choice.

Theorem 5.1. *Let $\alpha \geq 12\exp(-100d)$ be given. Consider the Bernoulli model (1.3) and the two parameter spaces $\mathcal{A}(a, k)$ and $\mathcal{A}(a, k_0)$, $1 \leq k_0 < k \leq m$. Furthermore assume that Assumption 1.1 is fulfilled, that $\sigma > 0$ is known, that $n \geq m\log(d)$ and consider the testing problem (5.1). Suppose*

$$\rho^2 \geq C\frac{m_1 m_2 k_0 d}{n} \asymp r_{n,m}^2(\mathcal{A}(a, k_0))$$

where $C = C(\alpha, a, U, \sigma) > 0$ is a constant. Then the test $\Psi_n := \mathbf{1}_{\{T_n > u_\alpha\}}$ where u_α is the quantile constant in (5.3) and T_n is as in (5.2) fulfills

$$\sup_{M \in \mathcal{A}(a,k_0)} \mathbb{E}_{M,\sigma}[\Psi_n] + \sup_{M \in \mathcal{A}(a,k),\ \|M - \mathcal{A}(a,k_0)\|_F^2 \geq \rho^2} \mathbb{E}_{M,\sigma}[1 - \Psi_n] \leq \alpha.$$

Now in order to apply Theorem 2.2 we use the soft-thresholding estimator proposed by Koltchinskii et. al. [29] which satisfies the oracle inequality (2.8) up to a log-factor in the trace regression model. That this holds in the Bernoulli-model as well with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1 - 1/d$ can be proven in a similar way and we sketch this in Proposition 8.3, removing the log-factor by using stronger bounds on the spectral norm of the stochastic term $(B_{ij}\epsilon_{ij})_{i,j}$.

This and Theorem 5.1 imply, using Theorem 2.2, that there exist honest and adaptive confidence sets in the Bernoulli model if the variance of the errors is known.

Corollary 5.1. *Let $\alpha \geq 2/d$ and $\alpha' \geq 12\exp(-100d)$ be given. Suppose that $\sigma > 0$ is known, that Assumption 1.1 is fulfilled and that $n \geq m\log(d)$. Then, for any $1 \leq k_0 < k \leq m$, there exists a honest confidence set C_n at the level α for the model $\mathcal{A}(a,k) \otimes \{\sigma\}$, i.e., for any $M_0 \in \mathcal{A}(a,k)$,*

$$\mathbb{P}_{M_0,\sigma}\left(M_0 \in C_n\right) \geq 1 - \alpha,$$

and C_n adapts to the sub-model $\mathcal{A}(a,k_0) \otimes \{\sigma\}$ at level α'.

5.2 The case of the Bernoulli model with unknown error variance.

In this subsection we assume again, as in Subsection 5.2, that the precise knowledge of the error variance σ is *not* available. Whereas in this case for the trace-regression model the construction of honest and adaptive confidence set was seen to be possible, we will now show that this is not the case for the Bernoulli model. We use again the duality between testing and confidence sets, this time applying Theorem 2.1. The next theorem gives a lower bound for the minimax rate of testing for the composite null hypothesis $H_0 : M \in \mathcal{A}(a,k_0)$ of M having rank at most k_0 against a rank-k alternative. To simplify the exposition we will consider only square matrices (but see the remark below) and also an asymptotic 'high-dimensional' framework where $\min(n,m) \to \infty$ and $k_0 = o(k)$. We formally allow for $k_0 = 0$, thus including the 'signal detection problem' when $H_0 : M = 0, \sigma^2 = 1$.

Theorem 5.2. *Suppose that Assumption 1.1 is satisfied for some $U \geq 2$ and assume $m = m_1 = m_2$. Furthermore, let $k = k_{n,m} \to \infty$ be such that $0 < k \leq m^{1/3}$ and $k^{1/4}\sqrt{m/n} < \min(1,a)/2$. For $0 \leq k_0 < k$ satisfying $k_0 = o(k)$ and a sequence $\rho = \rho_{n,m} \in (0,1/2)$ consider the testing problem*

$$H_0 : M \in \mathcal{A}(a,k_0),\ \sigma^2 = 1 \quad vs \quad H_1 : M \in \mathcal{A}(a,k),\ \|M - \mathcal{A}(a,k_0)\|_F^2 \geq m^2\rho^2,\ \sigma^2 = 1 - 4\rho^2. \tag{5.4}$$

If as $\min(n,m) \to \infty$,

$$\rho^2 = o\Big(\frac{\sqrt{k}m}{n}\Big), \tag{5.5}$$

then for any test Ψ we have that

$$\liminf_{\min(n,m)\to\infty}\left[\sup_{M\in\mathcal{A}(a,k_0)}\mathbb{E}_{M,1}[\Psi] + \sup_{M\in\mathcal{A}(a,k),\ \|M-\mathcal{A}(a,k_0)\|_F^2\geq m^2\rho^2}\mathbb{E}_{M,\sqrt{1-4\rho^2}}[1-\Psi]\right] \geq 1. \tag{5.6}$$

In particular, if $\Sigma \subset (0,U]$ contains the interval $[\sqrt{1-4\tau},1]$ where $\tau = \limsup_{n,m} k^{1/4}\sqrt{m/n}$, then (2.6) holds for the choices $\mathcal{A}_0 = \mathcal{A}(a,k_0)\otimes\Sigma, \mathcal{A} = \mathcal{A}(a,k)\otimes\Sigma$ and $\beta_0 = 1, \rho^* = \rho$.

Using Theorem 2.1 this implies the non-existence of honest and adaptive confidence sets in the model (1.3) if the variance of the errors is unknown and $k_0 = o(\sqrt{k})$. In particular adaptation to a constant rank k_0, $k_0 = O(1)$, is never possible if $k \to \infty$ as $\min(m,n) \to \infty$.

Corollary 5.2. *Assume that the conditions of Theorem 5.2 are fulfilled and that $k_0 = o(\sqrt{k})$. Then for any $\alpha, \alpha' > 0$ satisfying $0 < 2\alpha + \alpha' < 1$ a honest confidence set for the model $\mathcal{A}(a,k)\otimes\Sigma$ at level α that adapts to the sub-model $\mathcal{A}(a,k_0)\otimes\Sigma$ at level α' does not exist. In fact if $\alpha < 1/3$, we have for every honest confidence set C_n for the model $\mathcal{A}(a,k)\otimes\Sigma$ at level α and constant $c = c(a,U,\alpha)$ that*

$$\sup_{(M_0,\sigma)\in\mathcal{A}(a,k_0)\otimes\Sigma}\mathbb{E}_{M_0,\sigma}|C_n|_F^2 \geq c\frac{m^3\sqrt{k}}{n}.$$

The above results are formulated for square matrices ($m_1 = m_2$) to keep the technicalities in the proof at a reasonable level. One can adapt the proof of Theorem 5.2 to obtain a lower bound of the order $\rho^2 \gtrsim \sqrt{km_1m_2}/n$ which likewise leads to non-existence results for adaptive confidence sets for non-square matrices in relevant asymptotic regimes of k_0, k, m_1, m_2.

6 Conclusions

We have investigated confidence sets in two matrix completion models: the Bernoulli model and the trace regression model. In the trace regression model the construction of adaptive confidence sets is possible, even if the variance is unknown. Contrary to this we have shown that the information theoretic structure in the Bernoulli model is different; in this case the construction of adaptive confidence sets is not possible if the variance is unknown.

One interpretation is that in practical applications (e.g. recommender systems such as Netflix [3]) one should incentivise users to perform multiple ratings, to justify the use of the trace regression model and the proposed U-statistic confidence set.

Our proof only shows that one can not adapt to general low rank hypotheses if the variance is unknown. This covers the key cases where $k_0 = 1$ or more generally $k_0 = o(\sqrt{k})$. It remains an interesting open (and difficult) question whether the lower bound ρ in Theorem 5.2 is tight, but the answer to this question does not affect the main conclusions of our results on the existence of adaptive confidence sets in matrix completion problems.

7 Proofs

7.1 Proof of Theorem 2.2

Proof. Let Ψ_n be a test that attains the rate ρ with error probabilities bounded by $\min(\alpha/2, \alpha')$ and let $L = L(\min(\alpha/2, \alpha'))$ be the corresponding constant in (2.5). Let $\hat{f}$ denote an estimator that satisfies the oracle inequality (2.8) with probability of at least $1 - \alpha/2$. Define a confidence set

$$C_n := \{f \in \mathcal{A} : \ \mathcal{D}(\hat{f}, f) \leq K\left(r_n(\mathcal{A})\Psi_n + r_n(\mathcal{A}_0)(1 - \Psi_n)\right)\}$$

where $K > 0$ is a constant to be chosen.
We first prove that C_n is adaptive: If $f \in \mathcal{A}\backslash\mathcal{A}_0$ there is nothing to prove, and if $f \in \mathcal{A}_0$ we have

$$\mathbb{P}_f^n\left(|C_n|_{\mathcal{D}} > Kr_n(\mathcal{A}_0)\right) = \mathbb{P}_f^n(\Psi_n = 1) \leq \alpha'.$$

For coverage we investigate three distinct cases and note that

$$\sup_{f \in \tilde{\mathcal{A}}} \mathbb{P}_f^n\left(\mathcal{D}(\hat{f}, f) > Cr_n(\tilde{\mathcal{A}})\right) \leq \alpha/2 \tag{7.1}$$

where $C > 0$ is as in (2.8) and where $\tilde{\mathcal{A}} \in \{\mathcal{A}_0, \mathcal{A}\}$. Hence $\hat{f}$ is, by the oracle inequality, an adaptive estimator.
Then for $f \in \mathcal{A}_0$, by (7.1)

$$\mathbb{P}_f^n(f \notin C_n) \leq \mathbb{P}_f^n\left(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A}_0)\right) \leq \alpha/2 \leq \alpha$$

for $K \geq C$.
If $f \in \mathcal{A}\backslash\mathcal{A}_0$ and $\mathcal{D}(f, \mathcal{A}_0) \geq L\rho_n$, then for $K \geq C$

$$\begin{aligned}\mathbb{P}_f^n(f \notin C_n) &= \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A}), \Psi_n = 1) + \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A}), \Psi_n = 0) \\ &\leq \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A})) + \mathbb{P}_f^n(\Psi_n = 0) \leq \alpha.\end{aligned}$$

If $f \notin \mathcal{A}\backslash\mathcal{A}_0$ but $\mathcal{D}(f, \mathcal{A}_0) < L\rho_n$, then by the oracle inequality and since $\rho_n \leq C' r_n(\mathcal{A}_0)$ we have with probability at least $1 - \alpha/2$ for such f that

$$\mathcal{D}(\hat{f}, f) \leq C(\mathcal{D}(f, \mathcal{A}_0) + r_n(\mathcal{A}_0)) \leq CL\rho_n + Cr_n(\mathcal{A}_0) \leq C(LC' + 1)r_n(\mathcal{A}_0).$$

Thus we still have

$$\mathbb{P}_f^n\left(f \notin C_n\right) = \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A}_0)) \leq \alpha/2 \leq \alpha$$

for $K \geq C(LC' + 1)$. ∎

7.2 Proof of Theorem 4.2

Proof. Recall that

$$\mathbb{E}_{M_0,\sigma}\Big(\hat{R}_N|N, N>0\Big) = \frac{\|\hat{M}-M_0\|_F^2}{m_1 m_2} =: r. \tag{7.2}$$

Thus using Markov's inequality we have for $N>0$ that

$$\begin{aligned}\mathbb{P}_{M_0,\sigma}\left(M_0 \notin C_n|N, N>0\right) &\le \mathbb{P}_{M_0,\sigma}\left(|\hat{R}_N - r| > z_{\alpha,N}|N, N>0\right)\\ &\le \frac{\mathrm{Var}_{M_0,\sigma}\Big(\hat{R}_N\Big|N, N>0\Big)}{z_{\alpha,N}^2}.\end{aligned} \tag{7.3}$$

Using equation (7.2) we compute

$$\begin{aligned}\mathrm{Var}_{M_0,\sigma}\Big(\hat{R}_N\Big|N, N>0\Big) &= \frac{1}{N}\mathbb{E}_{M_0,\sigma}\Big(\big((Z_k - \langle\hat{M},\tilde{X}_k\rangle)(Z_k' - \langle\hat{M},\tilde{X}_i\rangle) - r\big)^2\Big|N, N>0\Big)\\ &\le \frac{1}{N}\left[\Big(\mathbb{E}\langle M_0-\hat{M}, X_1\rangle^4\Big) + 2\sigma^2 r + \sigma^4\right]\\ &= \frac{1}{N}\left[\frac{\|\hat{M}-M_0\|_{L^4}^4}{m_1 m_2} + 2\sigma^2 r + \sigma^4\right]\\ &\le \frac{U^4 + 8U^2a^2 + 16a^4}{N} = \alpha z_{\alpha,N}^2\end{aligned}$$

since $\|\hat{M}-M_0\|_\infty \le 2a$ and where we define $\|\hat{M}-M_0\|_{L^4}^4 := \sum_{i,j}(\hat{M}_{ij}-M_{ij})^4$. Hence (7.3) implies

$$\mathbb{P}_{M_0,\sigma}\left(M_0 \notin C_n|N>0\right) \le \alpha.$$

Moreover, as $\|\hat{M}-M_0\|_\infty \le 2a$ and $z_{\alpha,0} = 4a^2$, we have that $\mathbb{P}\left(M_0 \notin C_n|N=0\right) = 0$.

□

7.3 Proof of Theorem 5.1

Proof. If $M \in \mathcal{A}(a,k_0)$, then by definition of the infimum and u_α we have

$$\mathbb{E}_{M,\sigma}[\Psi] = \mathbb{P}_{M,\sigma}\left(T_n > u_\alpha\right) \le \mathbb{P}_\sigma\left(\frac{1}{\sqrt{2n}}\left|\sum_{ij} B_{ij}(\epsilon_{ij}^2 - \sigma^2)\right| > u_\alpha\right) \le \alpha/3.$$

The case $M \in \mathcal{A}(a,k), \|M - \mathcal{A}(a,k_0)\|_F^2 \ge \rho^2$ requires more elaborate arguments. Let A^* be a minimizer in (5.2). Then

$$\begin{aligned}\mathbb{E}_{M,\sigma}[1-\Psi] &= \mathbb{P}_{M,\sigma}\left(T_n < u_\alpha\right)\\ &= \mathbb{P}_\sigma\left(\left|\sum_{ij} B_{ij}[(A_{ij}^*-M_{ij})^2 - 2\epsilon_{ij}(A_{ij}^*-M_{ij}) + (\epsilon_{ij}^2-\sigma^2)]\right| < \sqrt{2n}u_\alpha\right).\end{aligned} \tag{7.4}$$

For $\rho \ge 8072a\sqrt{k_0 d/p} = 8072a\sqrt{m_1 m_2 k_0 d/n}$ we can apply Lemma 8.1 which yields a weaker version of the Restricted Isometry Property (RIP). Namely, Lemma 8.1 implies that the event

$$\Xi := \left\{\sum_{i,j} B_{ij}(A_{ij}-M_{ij})^2 \ge \frac{p}{2}\|A-M\|_F^2 \quad \forall A \in \mathcal{A}(a,k_0)\right\}, \quad M \in H_1,$$

occurs with probability of at least $1-2\exp(-100d)$. We can thus bound (7.4) by

$$\mathbb{P}_\sigma\left(\sup_{A\in\mathcal{A}(a,k_0)}\left[2\left|\sum_{i,j}B_{ij}\epsilon_{ij}(A_{ij}-M_{ij})\right|-\frac{\sum_{i,j}B_{ij}(A_{ij}-M_{ij})^2}{2}\right]>-\sqrt{n}u_\alpha,\Xi\right) \tag{7.5}$$

$$+\mathbb{P}_\sigma\left(\left|\sum_{i,j}B_{ij}(\epsilon_{ij}^2-\sigma^2)\right|>\frac{\sum_{i,j}B_{ij}(A^*_{ij}-M_{ij})^2}{2}-\sqrt{n}u_\alpha,\Xi\right)+2\exp(-100d). \tag{7.6}$$

The stochastic term (7.6) can be bounded using $d^2\geq 3n$ and that ρ is large enough. Indeed, on the event Ξ we have that

$$\frac{\sum_{i,j}B_{ij}(A^*_{ij}-M_{ij})^2}{2}\geq p\rho^2/4\geq(1+\sqrt{2})/\sqrt{3}du_\alpha\geq(1+\sqrt{2})\sqrt{n}u_\alpha$$

for $\rho\geq 2\sqrt{u_\alpha d/p}$ which implies together with the definition of u_α in (5.3) that (7.6) can be bounded by $\alpha/3+2\exp(-100d)$. For the cross term (7.5) we use the two following inequalities which, just as before, hold on the event $\Xi\ \forall\ A\in\mathcal{A}(a,k_0)$

$$\frac{\sum_{i,j}B_{ij}(A_{ij}-M_{ij})^2}{4}\geq\sqrt{n}u_\alpha\quad\text{and}\quad\frac{\sum_{i,j}B_{ij}(A_{ij}-M_{ij})^2}{8}\geq\frac{p\|A-M\|_F^2}{16}.$$

Hence, using also a peeling argument, (7.5) can be bounded by

$$\begin{aligned}
&\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty}\mathbb{P}_\sigma\left(\sup_{A\in\mathcal{A}(a,k_0),\ 2^s\leq p\|A-M\|_F^2\leq 2^{s+1}}\frac{\left|\sum_{i,j}B_{ij}\epsilon_{ij}(A_{ij}-M_{ij})\right|}{p\|A-M\|_F^2}>\frac{1}{16}\right)\\
\leq&\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty}\mathbb{P}_\sigma\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2^{s+1}}\left|\sum_{i,j}B_{ij}\epsilon_{ij}(A_{ij}-M_{ij})\right|>\frac{2^s}{16}\right)\\
=&\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty}\mathbb{P}_\sigma\left(Z(s)>\frac{2^s}{16}\right)
\end{aligned}\tag{7.7}$$

where we set the corresponding probability to 0 if the supremum is taken over an empty set and where we define

$$Z(s):=\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2^{s+1}}\left|\sum_{i,j}B_{ij}\epsilon_{ij}(A_{ij}-M_{ij})\right|.$$

Lemma 8.2 (with choices $z=16^2$, $\xi_{ij}=\epsilon_{ij}$, $t=2^s$ and $q=1$ there) implies for $\rho\geq 16144U\sqrt{k_0d/p}$ and for $2^s\geq p\rho^2/2$ that

$$\mathbb{P}_\sigma\left(Z(s)>\frac{2^s}{16}\right)\leq\exp\left(\frac{-2^s}{2097152U^2+517120aU}\right)$$

Hence, (7.7) can be upper bounded by

$$\begin{aligned}
\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty}\exp\left(\frac{-2^s}{2097152U^2+517120aU}\right)\leq&2\exp\left(-\frac{p\rho^2}{2097152U^2+517120aU}\right)\\
\leq&2\exp(-100d)
\end{aligned}\tag{7.8}$$

for $\rho\geq 16169U(a\vee U)\sqrt{d/p}$. Consequently (7.4) can be bounded by $\alpha/3+4\exp(-100d)\leq 2\alpha/3$ since $\alpha\geq 12\exp(-100d)$. □

7.4 Proof of Theorem 5.2

Proof. **Step I : Reduction to an easier testing problem between two distributions**
Assume without loss of generality that m is divisible by k. Suppose

$$\rho = \rho_{n,m} = \frac{vk^{1/4}\sqrt{m}}{\sqrt{n}} \tag{7.9}$$

where $v = v_{n,m}$ is a sequence such that $v = o(1)$, and assume w.l.o.g. that $0 < v \leq 1$. Moreover we denote $u = 2\rho$. For $1 \leq i \leq m,\ 1 \leq \kappa \leq k,\ 1 \leq j \leq m$ let

$$B_{ij} \overset{i.i.d.}{\sim} \mathcal{B}(p) \quad \text{and} \quad U_i^{\kappa} \overset{i.i.d.}{\sim} \mathcal{R} \quad \text{and} \quad V_j \overset{i.i.d.}{\sim} \mathcal{R},$$

where $\mathcal{B}(p)$ is a Bernoulli distribution of parameter $p = n/m^2$ and $\mathcal{R}$ is the standard Rademacher distribution $\Pr(V_1 = \pm 1) = 1/2$. Let $\mathcal{P}$ be a uniform random partition of $\{1, \ldots, m\}$ in k groups of size m/k, and denote by K_j, $K_j \in \{1, ..., k\}$, the label of element j of $\mathcal{P}$. Consider the following testing problem:

$$H_0' : M = 0 \quad \text{and} \quad \epsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{R}$$

against

$$\begin{gathered} H_1' : M_{ij} = uU_i^{K_j}V_j \\ \textit{and} \quad \epsilon_{ij} \sim \delta_{\{1-M_{ij}\}}(1+M_{ij})/2 + \delta_{\{-1-M_{ij}\}}(1-M_{i,j})/2 \end{gathered} \tag{7.10}$$

Note that the variance of ϵ_{ij} under H_0 is 1 and the variance of the noise under H_1 is

$$(1-M_{ij})^2(1+M_{ij})/2 + (-1-M_{ij})^2(1-M_{ij})/2 = (1-M_{ij})(1+M_{ij}) = 1-4\rho^2,$$

so the noise variables are homoscedastic across the (i,j)'s and $|\epsilon_{ij}| \leq 2 \leq U$. Let π be the distribution of M under H_1' and write ν_0 and ν_1 for the distribution of Y under H_0' and H_1', respectively.
Since the prior M in (7.10) consists of k i.i.d. scaled Rademacher vectors that each form m/k columns of M we have $\text{rank}(M) \leq k$ and $\|M\|_\infty = u = 2\rho \leq a$ for v small enough and since $k^{1/4}\sqrt{m/n} \leq a/2$. Thus $M \in \mathcal{A}(a,k)$. Then, reordering the columns of M we have

$$\|M - \mathcal{A}(a,k_0)\|_F^2 = \|M_{ord} - \mathcal{A}(a,k_0)\|_F^2$$

where M_{ord} is a $m \times m$ matrix with the $(((i-1)m/k)+1)$-th to the (im/k)-th columns each given by ur_i where r_i are $i.i.d$ Rademacher vectors of length m, $i = 1, ..., k$. Then (as in the proof of Theorem 1 in [16]) we transform M_{ord} into the $m \times k$ matrix $M_{ord}P = u\sqrt{m/k}R$ consisting of k column vectors $u\sqrt{m/k}r_i,\ i = 1, ..., k$. The $m \times k$ projection matrix P consists of k column vectors, the i-th having zero entries except for the indices $s \in [((i-1)m/k)+1, \ldots, im/k]$ where it equals $\sqrt{k/m}$. Hence P is an orthonormal projection matrix and we obtain

$$\|M - \mathcal{A}(a,k_0)\|_F^2 \geq \|(M_{ord} - \mathcal{A}(a,k_0))P\|_F^2 = \|u\sqrt{m/k}R - \mathcal{A}(a\sqrt{m/k},k,k_0)\|_F^2$$

where we define

$$\mathcal{A}(a,k,k_0) := \{A \in \mathbb{R}^{m\times k} :\ \|A\|_\infty \leq a \ \text{ and } \ \text{rank}(A) \leq k_0\}.$$

Therefore, if $\sigma_{min}(A)$ denotes the minimal singular value of a matrix A, we have that

$$\begin{aligned} \|M - \mathcal{A}(a,k_0)\|_F^2 &\geq \frac{m^2}{k}\|uR/\sqrt{m} - \mathcal{A}(a/\sqrt{m},k,k_0)\|_F^2 \\ &\geq \frac{m^2u^2}{k}(k-k_0)(\sigma_{min}(R/\sqrt{m}))^2 \\ &\geq \frac{m^2u^2}{2}(\sigma_{min}(R/\sqrt{m}))^2 \geq \frac{m^2u^2}{4} = m^2\rho^2 \end{aligned} \tag{7.11}$$

with probability going to 1, where we have used that $k - k_0 \geq k/2$ for m large enough (recall $k_0 = o(k)$) as well as the variational characterisation of minimal eigenvalues combined with Corollary 1 in [33] (with choices $n = m,\ p = k_1 = k,\ \theta = 0$ and $\Lambda_{min} = 1$ there) to lower bound $\sigma^2_{min}(R/\sqrt{m})$ by $1/2$.
To conclude, π is concentrated on H_1 and the primed testing problem above is, asymptotically, strictly easier than the testing problem (5.4) since H_0' is contained in H_0 and H_1' is asymptotically contained in H_1. Thus, we have for any test Ψ by a standard lower bound (as, e.g., in (6.23) in [20]) that for all $\eta > 0$

$$\mathbb{E}_{H_0}\Psi + \sup_{H_1}\mathbb{E}_{H_1}(1-\Psi) \geq \mathbb{E}_{H_0'}\Psi + \mathbb{E}_{H_1'}(1-\Psi) - o(1) \geq (1-\eta)\left(1 - \frac{d_{\chi^2}(\nu_0,\nu_1)}{\eta}\right) - o(1),$$

where $d_{\chi^2}(\nu_0,\nu_1)$ denotes the χ^2-distance between ν_0 and ν_1, which remains to be bounded.

Step II : Expectation over censored data
We define $I = [m] \times [m]$ and observe that the likelihood of the data under ν_0 is

$$L(Y_1, ...Y_{m,m}) = \prod_{(i,j)\in I}\Big((1-p)\mathbf{1}_{\{Y_{ij}=0\}} + \frac{p}{2}\mathbf{1}_{\{Y_{ij}=1\}} + \frac{p}{2}\mathbf{1}_{\{Y_{ij}=-1\}}\Big)$$

and that the likelihood of the data under ν_1 is

$$L(Y_1, ...Y_{m,m}) = \mathbb{E}_{M\sim\pi}\prod_{(i,j)\in I}\Big((1-p)\mathbf{1}_{\{Y_{ij}=0\}} + p(1/2+M_{ij}/2)\mathbf{1}_{\{Y_{ij}=1\}} + p(1/2-M_{ij}/2)\mathbf{1}_{\{Y_{ij}=-1\}}\Big).$$

Thus, the likelihood ratio $\mathcal{L}$ between these two distributions is given by

$$\mathcal{L} = \mathbb{E}_{M\sim\pi}\prod_{(i,j)\in I}\Big(\mathbf{1}_{\{Y_{ij}=0\}} + (1+M_{ij})\mathbf{1}_{\{Y_{ij}=1\}} + (1-M_{ij})\mathbf{1}_{\{Y_{ij}=-1\}}\Big).$$

So we have that

$$\begin{aligned}
d_{\chi^2}(\nu_0,\nu_1)^2 + 1 &= \mathbb{E}_{Y\sim\nu_0}\mathcal{L}^2 \\
&= \mathbb{E}_{Y\sim\nu_0}\Big[\mathbb{E}_{M\sim\pi}\prod_{(i,j)\in I}\Big(\mathbf{1}_{\{Y_{ij}=0\}} + (1+M_{ij})\mathbf{1}_{\{Y_{ij}=1\}} + (1-M_{ij})\mathbf{1}_{\{Y_{ij}=-1\}}\Big)\Big]^2 \\
&= \mathbb{E}_{M,M'\sim\pi}\prod_{i,j}\Big[\Big(1-p+\frac{p}{2}(1+M_{ij})(1+M'_{ij}) + \frac{p}{2}(1-M_{ij})(1-M'_{ij})\Big)\Big] \\
&= \mathbb{E}_{M,M'\sim\pi}\prod_{i,j}\Big[1+pM_{ij}M'_{ij}\Big]. \qquad (7.12)
\end{aligned}$$

where M' is an independent copy of M.

Step III : Conditioning over the cross information
Let $N_{r,r'}$ be the number of times where the couple $K_j = r, K'_j = r'$ occurs. That is,

$$N_{r,r'} := \sum_{j=1}^{m}\mathbf{1}_{\{K_j=r, K'_j=r'\}}.$$

We enumerate the elements inside these groups from 1 to $N_{r,r'}$. We write $\tilde{V}_j^{r,r'}$ for the corresponding enumeration of the V_j. Setting $\mathbf{N} = (N_{r,r'})_{r,r'}$ and using the definition of the prior, we compute

$$\begin{aligned}
\mathbb{E}_{M,M'\sim\pi}\prod_{i,j}\Big[1+pM_{ij}M'_{ij}\Big] &= \mathbb{E}_{\mathbf{N},U,\tilde{V},U',\tilde{V}'}\prod_{i=1}^{m}\prod_{r,r'\in\{1,...,k\}^2}\prod_{j=1}^{N_{r,r'}}\Big[1+pu^2U_i^r\tilde{V}_j^{r,r'}(U_i^{r'})'(\tilde{V}_j^{r,r'})'\Big] \\
&=: \mathbb{E}_{\mathbf{N}}\prod_{r,r'\in\{1,...,k\}^2}\mathcal{I}(N_{r,r'}) \qquad (7.13)
\end{aligned}$$

where we define for any $N = N_{r,r'} > 0$

$$\mathcal{I}(N) = \mathbb{E}_{X,W,X',W'} \prod_{i=1}^{m} \prod_{j=1}^{N} \Big[1 + pu^2 X_i W_j X_i' W_j'\Big]$$

and where $(X_i)_{i\le m}, (X_i')_{i\le m}, (W_i)_{j\le N}, (W_i')_{j\le N}$ are *i.i.d.* Rademacher random variables. Moreover, we set $\mathcal{I}_{r,r'}(0) = 0$.

Step IV : Bound on $\mathbb{E}_{\mathbf{N}} \prod_{r,r'\in\{1,\dots,k\}^2} \mathcal{I}(N_{r,r'})$.
In order to bound $\mathcal{I}(N)$ we use the following lemma proved below

Lemma 7.1. *Let $N = N_{r,r'}$. There exist constants $C_1, C_2, C_3 > 0$ such that for v small enough*

$$\mathcal{I}(N) \le \exp\Big(C_1 v^4 N/m\Big) \exp\Big(\frac{C_2 v^4 k^2 N}{m^2}\Big) \exp\Big(C_3 v^4 N^2 k^2/m^2\Big). \tag{7.14}$$

Using (7.12), (7.13) and (7.14) we have that

$$\begin{aligned}
&d_{\chi^2}(\nu_0,\nu_1)^2 + 1 \\
&= \mathbb{E}_{\mathbf{N}} \prod_{r,r'\in\{1,\dots,k\}^2} \mathcal{I}(N_{r,r'}) \qquad (7.15)\\
&\le \mathbb{E}_{\mathbf{N}} \left[\left(\exp\left(\frac{C_2 v^4 k^2}{m^2} \sum_{r,r'} N_{r,r'} \right) \right) \left(\exp\left(\frac{C_1 v^4}{m} \sum_{r,r'} N_{r,r'} \right) \right) \left(\prod_{r,r'\in\{1,\dots,k\}^2} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big) \right) \right] \\
&= \exp\left(C_2 v^4 \frac{k^2}{m} + C_1 v^4 \right) \mathbb{E}_{\mathbf{N}} \left[\prod_{r,r'\in\{1,\dots,k\}^2} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big) \right], \qquad (7.16)
\end{aligned}$$

since $\sum_{r,r'} N_{r,r'} = m$. We bound the expectation of the stochastic term in (7.16) using the following lemma proved below:

Lemma 7.2. *There exists a constant $C' > 0$ such that for v small enough we have*

$$\mathbb{E}_N \Big[\prod_{r,r'} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big) \Big] \le 1 + 2C' v^4 + \exp\Big(-m/k^2\Big). \tag{7.17}$$

Inserting (7.17) into (7.16) and summarizing all the steps we obtain

$$0 \le d_{\chi^2}(\nu_0,\nu_1)^2 \le C\left(v^2 + \exp\big(-m/k^2\big)\right) = o(1)$$

for a constant $C > 0$ and therefore, letting $\eta \to 0$,

$$\mathbb{E}_0[\Psi] + \sup_{H_1} \mathbb{E}_{H_1}[1-\Psi] \ge (1-\eta)\left(1 - \frac{d_{\chi^2}(\nu_0,\nu_1)}{\eta}\right) - o(1) = 1 - o(1).$$

□

Proof of Lemma 7.1. Note that, by construction of $\mathcal{P}$, we have that

$$N = N_{r,r'} \le m/k$$

since the number of j where $M_{\cdot,j}$ corresponds to $K_j = r$ is bounded by m/k. As the product of two independent Rademacher random variables is again a Rademacher random variable, we have

$$\mathcal{I}(N) = \mathbb{E}_{R,R'} \prod_{i=1}^{m} \prod_{j=1}^{N} \Big[1 + pu^2 R_i R_j'\Big],$$

where $R=(R_i)_{i=1}^m, R'=(R'_i)_{i=1}^N$ are independent Rademacher vectors of length m and N, respectively. The usual strategy to use $1+x\le e^x$ and then to bound iterated exponential moments of Rademacher variables (as in the proof of Theorem 1 of [16]) only works when $k=const$, and a more refined estimate is required for growing k, as relevant here.
We now bound $\mathcal{I}(N)$ for a fixed $N, m/k\ge N>0$. Using the binomial theorem twice we have

$$\begin{aligned}
\mathcal{I}(N) &= \mathbb{E}_{R'}\left[\Big[\frac{1}{2}\prod_{j=1}^N\big[1+pu^2R'_j\big]+\frac{1}{2}\prod_{j=1}^N\big[1-pu^2R'_j\big]\Big]^m\right]\\
&= \frac{1}{2^m}\sum_{s=1}^m\binom{m}{s}\Big[\frac{1}{2}\big[1+pu^2\big]^s\big[1-pu^2\big]^{m-s}+\frac{1}{2}\big[1-pu^2\big]^s\big[1+pu^2\big]^{m-s}\Big]^N\\
&= \frac{1}{2^m2^N}\sum_{s=1}^m\binom{m}{s}\sum_{q=1}^N\binom{N}{q}\big[1+pu^2\big]^{sq+(m-s)(N-q)}\big[1-pu^2\big]^{(m-s)q+s(N-q)}\\
&= \mathbb{E}_{Q,S}\Big[\big[1+pu^2\big]^{SQ+(m-S)(N-Q)}\big[1-pu^2\big]^{(m-S)Q+S(N-Q)}\Big]
\end{aligned}$$

with independent Binomial random variables $S\sim\mathcal{B}(1/2,m), Q\sim\mathcal{B}(1/2,N)$. If $A:=\frac{1-pu^2}{1+pu^2}$, we obtain

$$\begin{aligned}
\mathcal{I}(N) &= \mathbb{E}_{Q,S}\Big[\big[1+pu^2\big]^{mN}\left[\frac{1-pu^2}{1+pu^2}\right]^{SN+mQ-2SQ}\Big]\\
&= \big[1+pu^2\big]^{mN}\mathbb{E}_Q\Big[A^{mQ}\mathbb{E}_SA^{S(N-2Q)}\Big]\\
&= \big[1+pu^2\big]^{mN}\mathbb{E}_Q\left[A^{mQ}2^{-m}\left(A^{(N-2Q)}+1\right)^m\right]\\
&= \big[1+pu^2\big]^{mN}\mathbb{E}_Q\left[A^{Nm/2}\left(\frac{1}{2}A^{(N/2-Q)}+\frac{1}{2}A^{(-N/2+Q)}\right)^m\right]\\
&= \big[1-p^2u^4\big]^{mN/2}\mathbb{E}_Q\left(\frac{1}{2}A^{Q-N/2}+\frac{1}{2}A^{N/2-Q}\right)^m.
\end{aligned}$$

Now, we denote $x:=pu^2=4vk^{1/2}/m\le 1/2$ for v small enough. Furthermore, we Taylor expand $\log(A)$ about 1 up to second order, i.e.

$$\log(A)=\log(1-x)-\log(1+x)=-2x-\frac{1}{2}\left(\frac{1}{\xi_1^2}-\frac{1}{\xi_2^2}\right)x^2=:-2x-c(x)x^2$$

for $\xi_1\in[1/2,1]$, $\xi_2\in[1,3/2]$ and where $c(x)\in[0,16/9]$ since $x\le 1/2$. Hence, using also the inequality $e^x\le 1+x+x^2/2+x^3/6+2x^4$ we deduce

$$\begin{aligned}
\mathcal{I}(N) &\le \exp\big[-mNx^2/2\big]\mathbb{E}_Q\Big[\frac{1}{2}\exp\big(-2x(Q-N/2)-c(x)(Q-N/2)x^2\big)\\
&\quad+\frac{1}{2}\exp\big(-2x(N/2-Q)-c(x)(N/2-Q)x^2\big)\Big]^m\\
&\le \exp\big[-mNx^2/2\big]\\
&\quad\cdot\mathbb{E}_Q\Bigg[\frac{1}{2}\Big(1-2x(Q-N/2)-c(x)(Q-N/2)x^2+(-2x(Q-N/2)-c(x)(Q-N/2)x^2)^2/2\\
&\quad+(-2x(Q-N/2)-c(x)(Q-N/2)x^2)^3/6+2(-2x(Q-N/2)-c(x)(Q-N/2)x^2)^4\Big)\\
&\quad+\frac{1}{2}\Big(1-2x(N/2-Q)-c(x)(N/2-Q)x^2+(-2x(N/2-Q)-c(x)(N/2-Q)x^2)^2/2\\
&\quad+(-2x(N/2-Q)-c(x)(N/2-Q)x^2)^3/6+2(-2x(N/2-Q)-c(x)(N/2-Q)x^2)^4\Big)\Bigg]^m.
\end{aligned}$$

Since $x \le 1/2$ and $|N/2 - Q|x \le 1/4$ there exist two constants $c_2 = c_2(x) = c(x)/2 + c(x)^2/32 \le 1$ and $c_1 = c_1(x) = 32 + 32c(x) + 12c(x)^2 + 2c(x)^3 + c(x)^4/8 \le 140$ such that the last equation above can be bounded by

$$\begin{aligned}
&\le \exp\big[-mNx^2/2\big]\mathbb{E}_Q\Big[1 + 2x^2(Q-N/2)^2 + c_1|Q-N/2|^4x^4 + c_2|Q-N/2|x^2\Big]^m \\
&\le \exp\big[-mNx^2/2\big]\mathbb{E}_Q \exp\Big[mx^2(N-2Q)^2/2 + c_1 m(Q-N/2)^4x^4 + c_2 m|Q-N/2|x^2\Big] \\
&= \mathbb{E}_Q\Big[\exp\Big(\frac{m}{2}\big(x^2(2Q-N)^2 - Nx^2\big)\Big)\exp\Big(c_1 m(Q-N/2)^4x^4 + c_2 m|Q-N/2|x^2\Big)\Big].
\end{aligned}$$

Using the Cauchy-Schwarz inequality twice, this implies that

$$\begin{aligned}
\mathcal{I}(N) \le \sqrt{\mathbb{E}_Q\Big[\exp\Big(mx^2N\big((2Q-N)^2/N - 1\big)\Big)\Big]}\Bigg[\mathbb{E}_Q\Big[\exp\Big(c_1 mx^4(N-2Q)^4/4\Big)\Big] \\
\cdot\, \mathbb{E}_Q\Big[\exp\Big(2c_2 m|2Q-N|x^2\Big)\Big]\Bigg]^{1/4} =: \sqrt{(I)}(II)^{1/4}(III)^{1/4}.
\end{aligned}$$

Step 1 : Bound on term (III)
Since $Q \sim \mathcal{B}(1/2, N)$, since $(2Q-N)$ is symmetric and since $2c_2mx^2 \le 1/2$ we have that

$$\begin{aligned}
(III) &= \mathbb{E}_Q\Big[\exp\Big(2c_2 m|2Q-N|x^2\Big)\Big] \le 2\mathbb{E}_Q\Big[\exp\Big(2c_2 m(2Q-N)x^2\Big)\Big] \\
&= 2\Big[\exp\Big(2c_2mx^2\Big) + \exp\Big(-2c_2mx^2\Big)\Big]^N \le 2\Big[1 + 8c_2^2m^2x^4\Big]^N \\
&\le \exp\big(8c_2^2m^2x^4N\big) \le \exp\big(\frac{C_2v^4k^2N}{m^2}\big). \qquad (7.18)
\end{aligned}$$

Step 2 : Term (II)
We use $mN^2x^4 \le 64v^4/m$, $(N-2Q)^2 \le N^2$ and $N \le m/k$ to obtain

$$(II) \le \mathbb{E}_Q\Big[\exp\Big(64c_1v^4N/m \cdot (N-2Q)^2/N\Big)\Big].$$

Since $Q \sim \mathcal{B}(1/2, N)$ the Rademacher average $Z = (N-2Q)/\sqrt{N}$ is sub-Gaussian with sub-Gaussian constant at most 1. It hence satisfies (e.g., equation (2.24) in [20]) for $c > 2$

$$\mathbb{E}\exp\{Z^2/c^2\} \le 1 + \frac{2}{c^2/4 - 1} \le e^{c_3c^{-2}},$$

which for v small enough and the choice $c^{-2} = 64c_1v^4N/m$ implies for some constant C_1 that

$$(II) \le \exp\left(\frac{4C_1v^4N}{m}\right).$$

Step 3 : Term (I)
We have that

$$\begin{aligned}
(I) &= \mathbb{E}_Q\Big[\exp\Big(mNx^2\left[\frac{(2Q-N)^2}{N} - 1\right]\Big)\Big] \\
&= \mathbb{E}\left[\exp\left(\frac{16v^2Nk}{m}\left[\frac{1}{N}\left(\sum_{i=1}^N \varepsilon_i\right)^2 - 1\right]\right)\right] = \mathbb{E}\left[\exp\left(\frac{16v^2k}{m}\sum_{i\ne j, i,j\le N}\varepsilon_i\varepsilon_j\right)\right],
\end{aligned}$$

where ε_i are *i.i.d.* Rademacher random variables. If $A = (a_{ij})$ is a symmetric matrix with all elements on the diagonal equal to zero, then for the Laplace transform of an order-two Rademacher chaos $Z = \sum_{i,j} a_{ij}\varepsilon_i\varepsilon_j$ we have the inequality

$$\mathbb{E}e^{\lambda Z} \le \exp\left\{\frac{16\lambda^2\|A\|_F^2}{2\left(1 - 64\|A\|\lambda\right)}\right\}, \quad \lambda > 0,$$

see, e.g., Exercise 6.9 on p.212 in [5] with $\mathcal{T} = \{A\}$. Now take $A = (\delta_{i\neq j})_{i,j\leq N}$ so that we have $\|A\| \leq N$ and for v small enough $16v^2kN/m \leq 16v^2 \leq 1/128$.

$$\mathbb{E}\Big[\exp\Big(\frac{16v^2k}{m}\sum_{i\neq j,i,j\leq N}\varepsilon_i\varepsilon_j\Big)\Big] \leq \exp\left(\frac{16^3v^4k^2\|A\|_F^2}{2m^2(1-1024v^2k\|A\|/m)}\right) \leq \exp\left(\frac{16^3v^4k^2N^2}{m^2}\right)$$

and therefore we conclude for a constant $C_3 > 0$ that

$$(I) \leq \exp\Big(2C_3v^4k^2N^2/m^2\Big). \tag{7.19}$$

Step 4 : Conclusion on $\mathcal{I}(N)$
Combining the bounds for (I), (II) and (III) with the bound on $\mathcal{I}(N)$ we have that

$$\mathcal{I}(N) \leq \exp\Big(C_2v^4k^2N/m^2\Big)\exp\Big(C_1v^4N/m\Big)\exp\Big(C_3v^4k^2N^2/m^2\Big).$$

□

Proof of Lemma 7.2. We bound the expectation by bounding it separately on two complementary events. For this we consider the event ξ where all $N_{r,r'}$ are upper bounded by $\tau := 15m/k^2$, assumed to be an integer (if not replace it by its integer part plus one in the argument below). More precisely we define

$$\xi = \Big\{\forall r \leq k, \forall r' \leq k \ : \ N_{r,r'} \leq \tau\Big\}.$$

Note that $\{N_{r,r'} > \tau\}$ occurs only if the size of the intersection of the class r of partition $\mathcal{P}$ with the class r' of partition $\mathcal{P}'$ is larger than τ. This means that at least τ elements among m/k elements of the class r', must belong to the class r. The positions of these τ elements can be taken arbitrarily within the m/k elements. For the first element, among those τ, the probability to belong to the class r is $\frac{m/k}{m}$. For the second element this probability is $\frac{m/k}{m-1}$ or $\frac{(m/k)-1}{m-1}$ and so on. All these probabilities are smaller than $(m/k)/(m-m/k+1)$. Therefore we have

$$\mathbb{P}_{\mathbf{N}}(N_{r,r'} > \tau) \leq \binom{m/k}{\tau}\left(\frac{m/k}{m-m/k+1}\right)^{\tau} \leq \frac{(m/k)^{\tau}}{\tau!}(2/k)^{\tau} \leq 2^{\tau}(m/k^2)^{\tau}\tau^{-\tau}e^{\tau} \leq e^{-\tau},$$

where we use $\binom{m/k}{\tau} \leq \frac{(m/k)^{\tau}}{\tau!}$ and Stirling's formula. Using a union bound this implies that the probability of ξ is lower bounded by $1 - k^2\exp(-15m/k^2)$.
We have on the event ξ

$$\begin{aligned}
&\mathbb{E}_{\mathbf{N}}\Big[\mathbf{1}\{\xi\}\prod_{r,r'\in\{1,\dots,k\}^2}\exp\Big(C_3v^4N_{r,r'}^2k^2/m^2\Big)\Big]\\
&\leq \exp\Big(C_3v^4k^2\cdot 15^2(m/k^2)^2k^2/m^2\Big)\Big]\\
&\leq \exp\Big(C'v^4\Big) \leq 1 + 2C'v^4.
\end{aligned}$$

for $C' = 225C_3$ and for v small enough. Moreover, by definition of $N_{r,r'}$, we have that $N_{r,r'} \leq m/k$ and $\sum_{r,r'}N_{r,r'} = m$. Hence

$$\sum_{r,r'}N_{r,r'}^2 \leq km^2/k^2 = m^2/k$$

which implies that on ξ^C

$$\begin{aligned}
&\mathbb{E}_{\mathbf{N}}\Big[\mathbf{1}\{\xi^C\}\prod_{r,r'\in\{1,\dots,k\}^2}\exp\Big(C_3v^4N_{r,r'}^2k^2/m^2\Big)\Big]\\
&\leq \mathbb{P}_{\mathbf{N}}(\xi^C)\exp\Big(C_3v^4k\Big)\\
&\leq k^2\exp\Big(-15m/k^2 + C_3v^4k\Big)\\
&\leq k^2\exp\Big(-3m/k^2\Big) \leq \exp\Big(-m/k^2\Big),
\end{aligned}$$

for v small enough and since $k^3 \leq m$. Thus, combining the bounds on ξ and ξ^C, we have that

$$\mathbb{E}_N\Big[\prod_{r,r'} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big)\Big] \leq 1 + 2C'v^4 + \exp\Big(-m/k^2\Big).$$

□

8 Auxiliary results

8.1 Proof of Lemma 4.1

Proof. Assume that among the first $n/4$ samples we have less than $n/8$ entries that are sampled twice - otherwise the result holds since $n/8 \geq n^2/64m_1m_2$ for $n \leq m_1m_2$. Then, among the first $n/4$ samples, there are at least $n/8$ distinct elements of $\mathcal{B}$, the set of all standard basis matrices in $\mathbb{R}^{m_1 \times m_2}$, that have been sampled at least once. We write $\mathcal{S}$ for the set of *distinct* elements of $\{X_i\}_{i \leq n/4}$ and obviously have $|\mathcal{S}| \geq n/8$. Hence, by definition of the sampling scheme, we have that

$$\mathbb{P}(X_i \in \mathcal{S}) \geq \frac{n}{8m_1m_2}, \qquad n/4 < i \leq n/2.$$

Furthermore, when sampling an element from $\mathcal{S}$ we have to remove this element from $\mathcal{S}$ as we have to use the entry that is stored in $\mathcal{S}$ to form a pair of entries. Hence the probability to sample another element from $\mathcal{S}$ decreases and is bounded by

$$\mathbb{P}(X_j \in \mathcal{S}\backslash\{X_i\} \big| X_i \in \mathcal{S}) \geq \frac{n-1}{8m_1m_2}$$

for $n/4 < i < j < n/2$. We deduce by induction for $j > i + k$ and $k \leq n/2 - i - 1$ that

$$\mathbb{P}(X_j \in \mathcal{S}\backslash\{X_i, ..., X_{i+k}\} \big| X_i, ..., X_{i+k} \in \mathcal{S}) \geq \frac{n-k}{8m_1m_2}$$

which yields

$$\begin{aligned}\mathbb{P}\left(N \geq \frac{n^2}{64m_1m_2}\right) &\geq \mathbb{P}\left(\sum_{n/4<i\leq n/2} \mathbf{1}_{\{X_i \in \mathcal{S}\}} \geq \frac{n^2}{64m_1m_2}\right) \\ &\geq \mathbb{P}\left(\sum_{n/4<i\leq n/2} \mathbf{Z}_i \geq \frac{n^2}{64m_1m_2}\right) \end{aligned} \tag{8.1}$$

where $\mathbf{Z}_i$ can be taken to be Bernoulli random variables with success probability

$$p' = \frac{n - \frac{n^2}{64m_1m_2}}{8m_1m_2}.$$

Then, Bernstein's inequality for bounded random variables (see e.g. Theorem 3.1.7 in [20]), (8.1) and the estimates

$$\mathbb{E}\left[\sum_{n/4<i\leq n/2} \mathbf{Z}_i\right] \geq \frac{n^2}{33m_1m_2}$$

which holds for $n \leq m_1m_2$ and

$$\mathrm{Var}\left(\sum_{n/4<i\leq n/2} \mathbf{Z}_i\right) \leq \frac{n^2}{32m_1m_2}$$

imply that

$$\mathbb{P}\left(N \geq \frac{n^2}{64m_1m_2}\right) \geq 1 - \mathbb{P}\left(\sum_{n/4<i\leq n/2} \mathbf{Z}_i - \mathbb{E}\left[\sum_{n/4<i\leq n/2} \mathbf{Z}_i\right] \leq \frac{-n^2}{72m_1m_2}\right) \geq 1 - \exp\left(\frac{n^2}{372m_1m_2}\right).$$

□

8.2 Lemma 8.1

Lemma 8.1. *Consider the Bernoulli model (1.3) and assume $n \geq m\log(d)$. Then, with probability at least $1-2\exp(-100d)$ we have for any given $M \in \mathcal{A}(a,m)$ that*

$$\sup_{A\in\mathcal{A}(a,m),\ \|M-A\|_F\geq Ca\sqrt{(\mathrm{rank}(A)\vee 1)d/p}} \left[\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| - \frac{p}{2}\|M_0-A\|_F^2\right] \leq 0$$

where $C = 8072$.

Proof. We have, using a union bound, that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\mathcal{A}(a,m),\ \|M-A\|_F\geq Ca\sqrt{(\mathrm{rank}(A)\vee 1)d/p}} \left[\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| - \frac{p}{2}\|M_0-A\|_F^2\right] > 0\right)\\
&\leq \sum_{k=1}^{m}\mathbb{P}\left(\sup_{A\in\mathcal{A}(a,k),\ p\|M-A\|_F^2\geq C^2a^2kd} \left[\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| - \frac{p}{2}\|A-M\|_F^2\right] > 0\right). \qquad (8.2)
\end{aligned}$$

Then, using a peeling argument each of the terms in (8.2) can be bounded by

$$\begin{aligned}
&\sum_{s\in\mathbb{N}:\ C^2a^2kd/2\leq 2^s<\infty}\mathbb{P}\left(\sup_{A\in\mathcal{A}(a,k),\ 2^s\leq p\|A-M\|_F^2\leq 2^{s+1}}\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| > 2^s/2\right)\\
&\leq \sum_{s\in\mathbb{N}:\ C^2a^2kd/2\leq 2^s<\infty}\mathbb{P}\left(\sup_{A\in\mathcal{A}(a,k),\ p\|A-M\|_F^2\leq 2^{s+1}}\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| > 2^s/2\right) \qquad (8.3)
\end{aligned}$$

with the convention that if the supremum is taken over an empty set the corresponding probability is set equal to 0. For the cases where the supremum is not taken over an empty set, we apply Lemma 8.2 (with choices $\xi_{ij}=1$, $q=2$, $z=4$, $U=1$ and $t=2^s$ there) and obtain for

$$Z(s) := \sup_{A\in\mathcal{A}(a,k),\ p\|A-M\|_F^2\leq 2^{s+1}}\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right|$$

that we can bound

$$\mathbb{P}\left(Z(s) > 2^s/2\right) \leq \exp\left(\frac{-2^s}{260352a^2}\right)$$

Hence, (8.3) can be upper bounded by

$$\sum_{s\in\mathbb{N}:\ Ca^2kd/2\leq 2^s<\infty}\exp\left(\frac{-2^s}{260352a^2}\right) \leq 2\exp\left(-\frac{C^2kd}{260352}\right) \leq 2\exp(-101d).$$

The result then follows by noting that $\log(m)\leq d$. ∎

8.3 Lemma 8.2

Lemma 8.2. *Consider the Bernoulli model (1.3). Suppose that ξ_{ij} are independent random variables with $\max_{ij}|\xi_{ij}| \leq U$ and that $m\log(d)\leq n$. Let $z>0$, $q\in\{1,2\}$, $M\in\mathcal{A}(a,m)$ and $1\leq k_0<m$ be given. Finally, for $C=1009$ suppose that $t\in\mathbb{R}_+$ is such that $t\geq C^2z(4a)^{2q-2}U^2k_0d/2$ and that the supremum in*

$$Z(t) := \sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2t}\left|\sum_{i,j}\left[(B_{ij}\xi_{ij}-\mathbb{E}B_{ij}\xi_{ij})(A_{ij}-M_{ij})^q\right]\right|$$

is not empty. Then,

$$\mathbb{P}\left(Z(t) > \frac{t}{\sqrt{z}}\right) \leq \exp\left(\frac{-t}{32^2(8(2a)^{2q-2}U^2z + 505(2a)^qU\sqrt{z}/32)}\right) \qquad (8.4)$$

Proof. We first bound $\mathbb{E}Z(t)$ and then apply Talagrand's [39] inequality. Using symmetrization (e.g. Theorem 3.1.21 in [20]) and two contraction inequalities (e.g. Theorems 3.1.17 and 3.2.1 in [20]), we obtain that

$$\begin{aligned}
\mathbb{E}Z(t) &\le 2U\mathbb{E}\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\le 2t}\left|\sum_{i,j}B_{ij}\varepsilon_{ij}(A_{ij}-M_{ij})^q\right|\right)\\
&\le 2(4a)^{q-1}U\mathbb{E}\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\le 2t}\left|\sum_{i,j}B_{ij}\varepsilon_{ij}(A_{ij}-M_{ij})\right|\right)\\
&\le 2(4a)^{q-1}U\mathbb{E}\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\le 2t}|\langle\Sigma_R,A-A_0\rangle|\right)+2(4a)^{q-1}U\mathbb{E}\,|\langle\Sigma_R,A_0-M\rangle|\\
&\le 8(4a)^{q-1}U\sqrt{k_0t/p}\mathbb{E}\,\|\Sigma_R\|+2(4a)^{q-1}U\mathbb{E}\,|\langle\Sigma_R,A_0-M\rangle|\,. \qquad (8.5)
\end{aligned}$$

where ε_{ij} are independent Rademacher random variables, $\Sigma_R := \left(B_{ij}\varepsilon_{ij}\right)_{ij}$ and where A_0 is an arbitrary element in $\mathcal{A}(a,k_0)$ such that $p\|A_0-M\|_F^2 \le 2t$. Such an A_0 exists as soon as the supremum is not taken over an empty set. An extension of Corollary 3.6 in [1] to rectangular matrices by self-adjoint dilation (e.g. section 3.1. in [1]) implies (with choices $\xi_{ij}=B_{ij}\varepsilon_{ij}/\sqrt{p}$, $b_{ij}=\sqrt{p}$, $\alpha=3$ and $\sigma=\max\left(\max_j\sqrt{\sum_{i=1}^{m_1}b_{ij}^2},\max_i\sqrt{\sum_{j=1}^{m_2}b_{ij}^2}\right)\le\sqrt{pd}$ there) that

$$\mathbb{E}\,\|\Sigma_R\|\le e^{2/3}(2\sqrt{pd}+42\sqrt{\log(d)})\le 86\sqrt{pd}$$

since $m\log(d)\le n$. For the second term in (8.5) we have

$$\begin{aligned}
\mathbb{E}|\langle\Sigma_R,A_0-M\rangle| &\le \left(\mathrm{Var}(\langle\Sigma_R,A_0-M\rangle)\right)^{1/2}\\
&= \left(p\|A_0-M\|_F^2\right)^{1/2}\le\sqrt{2t}.
\end{aligned}$$

Hence, for $C^2z(4a)^{2q-2}U^2k_0d/2\le t$ and since $C=1009$ we have that

$$\mathbb{E}Z(t)\le 688(4a)^{q-1}U\sqrt{k_0td}+2(4a)^{q-1}U\sqrt{2t}\le 31t/(32\sqrt{z}). \qquad (8.6)$$

We now make use of the following inequality due to Talagrand [39], which in the current form with explicit constants can be obtained by inverting the tail bound in Theorem 3.3.16 in [20].

Theorem 8.1. *Let $(S,\mathcal{S})$ be a measurable space and let $n\in\mathbb{N}$. Let X_k, $k=1,\dots,n$ be independent S-valued random variables and let $\mathcal{F}$ be a countable set of functions $f=(f_1,...,f_n):S^n\to[-K,K]^n$ such that $\mathbb{E}f_k(X_k)=0$ for all $f\in\mathcal{F}$ and $k=1,...,n$. Set*

$$Z:=\sup_{f\in\mathcal{F}}\sum_{k=1}^n f_k(X_k).$$

Define the variance proxy

$$V_n:=2K\mathbb{E}Z+\sup_{f\in\mathcal{F}}\sum_{k=1}^n\mathbb{E}\left[(f_k(X_k))^2\right].$$

Then, for all $t\ge 0$,

$$\mathbb{P}\left(Z-\mathbb{E}Z\ge t\right)\le\exp\left(\frac{-t^2}{4V_n+(9/2)Kt}\right).$$

The functional $A\to\|A-M\|_F^2$ is continuous on the compact set of matrices $\{A\in\mathcal{A}(a,k_0):\|A-M\|_F^2\le 2t\}$, hence by continuity and compactness the supremum is attained over a countable subset. Thus we may apply Talagrand's inequality to $Z(t)$. We have for our particular case, since $\sup_{f\in\mathcal{F}}|f(X)|=\sup_{f\in\{\mathcal{F}\bigcup\{-\mathcal{F}\}}f(x)$, that

$$X_{ij}=B_{ij}\xi_{ij}-\mathbb{E}B_{ij}\xi_{ij},\quad S=[-2U,2U]$$

$$\mathcal{F} = \Big\{ f : S^{m_1 \times m_2} \to [-2(2a)^q U, 2(2a)^q U]^{m_1 \times m_2},\ f_{ij}(X_{ij}) = (-1)^l X_{ij}(A_{ij} - M_{ij})^q,$$
$$A \in \mathcal{A}(a, k_0),\ p\|A - M\|_F^2 \le 2t,\ l \in \{0, 1\} \Big\}$$

and moreover

$$\begin{aligned}
&\sup_{(A,l),\ A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\le 2t,\ l\in\{0,1\}} \sum_{i,j} \mathbb{E}\left[\left((-1)^l(B_{ij}\xi_{ij} - \mathbb{E}B_{ij}\xi_{ij})(A_{ij} - M_{ij})^q\right)^2\right] \\
\le &(2a)^{2q-2} \sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\le 2t} \sum_{i,j} \mathrm{Var}(B_{ij}\xi_{ij})(A_{ij} - M_{ij})^2 \\
\le &(2a)^{2q-2}U^2 \sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\le 2t} \sum_{i,j} p(A_{ij} - M_{ij})^2 \le 2(2a)^{2q-2}U^2 t.
\end{aligned}$$

Therefore, using our previous estimate in (8.6) for $\mathbb{E}Z(t)$ as well, we have for the variance proxy $V_{m_1 m_2}$ that

$$V_{m_1 m_2} \le 2(2a)^{2q-2}U^2 t + 31(2a)^q U t/(8\sqrt{z}).$$

Hence, using (8.6) and Talagrand's inequality, we obtain

$$\mathbb{P}\left(Z(t) > \frac{t}{\sqrt{z}}\right) \le \mathbb{P}\left(Z(t) - \mathbb{E}Z(t) > \frac{t}{32\sqrt{z}}\right) \le \exp\left(\frac{-t}{32^2(8(2a)^{2q-2}U^2 z + 505(2a)^q U\sqrt{z}/32)}\right).$$

□

8.4 An oracle estimator in the Bernoulli model

Here we prove that the soft-thresholding estimator proposed by Koltchinskii et. al. [29] for the trace-regression setting fulfills the oracle inequality (2.8) in the Bernoulli model.
Their estimator is defined as

$$\hat{M} \in \operatorname*{arg\,min}_{A\in\mathbb{R}^{m_1\times m_2}} \left(\frac{\|A\|_F^2}{m_1 m_2} - \frac{2}{n}\langle Y, A\rangle + \lambda\|A\|_*\right) \tag{8.7}$$

where λ is a tuning parameter which we choose as

$$\lambda = 3\left(\frac{3\sqrt{2}\sigma + \sqrt{2C}U}{\sqrt{mn}}\right) \tag{8.8}$$

where $C > 0$ is the constant in Corollary 3.12 in [1].

Proposition 8.3. *Consider the Bernoulli model (1.3). Assume $n \ge m\log(d)$ and that Assumption 1.1 is fulfilled. Let $\hat{M}$ be given as in (8.7) with a choice of λ as in (8.8). Then, with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1 - 1/d$ we have for any $M_0 \in \mathcal{A}(a, m)$ that*

$$\begin{aligned}
\frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} &\le \inf_{A\in\mathbb{R}^{m_1\times m_2}} \left(\frac{\|M_0 - A\|_F^2}{m_1 m_2} + C\frac{\mathrm{drank}(A)}{n}\right) \\
&\le \inf_{k\in\{0,\dots,m\}} \left(\frac{\|M_0 - \mathcal{A}(a,k)\|_F^2}{m_1 m_2} + C\frac{dk}{n}\right)
\end{aligned}$$

for a constant $C = C(a, \sigma, U) > 0$.

Proof. Going through the proof of Theorem 2 and Corollary 2 in [29] line by line we see that we only need to bound the spectral norm of the matrix

$$\Sigma := \frac{1}{n}(B_{ij}\epsilon_{ij})_{i,j}$$

by $\lambda/3$ with high probability. Using self-adjoint dilation to generalize Corollary 3.12 and Remark 3.13 in [1] for rectangular matrices (with choices $\varepsilon = 1/2$, $\tilde{\sigma_*} = U$ and

$$\tilde{\sigma} = \max\left(\max_j \sqrt{\sum_{i=1}^{m_1} \mathbb{E}_\sigma B_{ij}^2 \epsilon_{ij}^2}, \max_i \sqrt{\sum_{j=1}^{m_2} \mathbb{E}_\sigma B_{ij}^2 \epsilon_{ij}^2}\right) = \sigma\sqrt{n/m}$$

there) we obtain

$$\mathbb{P}_\sigma\left(\left\|\sum_{i=1}^n \varepsilon_i X_i\right\| > 3\sqrt{2}\sigma\sqrt{\frac{n}{m}} + t\right) \leq d\exp\left(-\frac{-t^2}{C_1 U^2}\right)$$

for a constant $C_1 > 0$. Choosing $t = \sqrt{2C_1}U\sqrt{\frac{n}{m}}$ and using that $n \geq m\log(d)$ yields that Ξ occurs with $\mathbb{P}_\sigma$-probability at least $1 - 1/d$. □

Acknowledgements

The work of A. Carpentier is supported by the DFGs Emmy Noether grant MuSyAD (CA 1488/1-1). The work of O. Klopp was conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The work of M. Löffler was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016516/1 and the European Research Council (ERC) grant No. 647812. The latter ERC grant also supported R. Nickl, who is further grateful to A. Tsybakov and ENSAE Paris for their hospitality during a visit in April 2016 where part of this research was undertaken.

References

[1] A. S. Bandeira and R. van Handel. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.* **44**(4), 2479-2506

[2] Y. Baraud. (2004). Confidence balls in Gaussian regression. *Ann. Statist.*, **32**(2):528-551.

[3] J. Bennett and S. Lanning. (2007). The netflix prize. *Proceedings of KDD Cup and Workshop.*

[4] P. Biswas, T. Liang, T. Wang, and Y. Ye. (2006). Semidefinite programming based algorithms for sensor network localization. *ACM Trans. Sen. Netw.*, **2**(2):188–220.

[5] S. Boucheron, G. Lugosi and P. Massart. (2013). *Concentration inequalities.* Oxford University Press.

[6] A. D. Bull and R. Nickl. (2013). Adaptive confidence sets in L^2. *Probab. Theory Related Fields*, **156**(3):889–919.

[7] T. T. Cai and M. G. Low. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.*, **32**(5):1805–1840.

[8] T. T. Cai and Z. Guo. (2016). Accuracy assessment for high-dimensional linear regression. http://arxiv.org/abs/1603.03474

[9] T. T. Cai and W. Zhou. (2016). Matrix completion via max-norm constrained optimization. *Electron. J. Statist.*, **10**(1):1493–1525.

[10] E. J. Candès and B. Recht. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**(6):717–772.

[11] E. J. Candès and T. Tao. (2010). The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, **56**(5):2053–2080.

[12] Candès, E. J. and Plan, Y. (2009). Matrix completion with noise. *Proceedings of IEEE*, **98(6)**, 925-936.

[13] E. J. Candès and Y. Plan. (2011). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory*, **57**(4):2342–2359.

[14] A. Carpentier. (2013). Honest and adaptive confidence sets in L^p. *Electron. J. Statist.*, **7**:2875–2923.

[15] A. Carpentier, J. Eisert, D. Gross, and R. Nickl. (2015). Uncertainty Quantification for Matrix Compressed Sensing and Quantum Tomography Problems. http://arxiv.org/abs/1504.03234

[16] A. Carpentier and R. Nickl. (2015). On signal detection and confidence sets for low rank inference problems. *Electron. J. Statist.*, **9**(2):2675–2688.

[17] S. Chatterjee. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.*, **43**(1):177–214.

[18] E. C. Chi, H. Zhou, G.K. Chen, D.O. Del Vecchyo and K. Lange. (2013). Genotype imputation via matrix completion. *Genome Res.*, **23**(3):509-518.

[19] E. Giné and R. Nickl. (2010). Confidence bands in density estimation. *Ann. Statist.*, **38**(2):1122–1170.

[20] E. Giné and R. Nickl. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Methods*. Cambridge University Press.

[21] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM*, **35**(12):61–70.

[22] D. Gross. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, **57**(3):1548–1566.

[23] M. Hoffmann and R. Nickl. (2011). On adaptive inference and confidence bands. *Ann. Statist.*, **39**(5):2383–2409.

[24] P. Jain, P. Netrapalli and S. Sanghavi. (2013). Low-rank matrix completion using alternating minimization. *STOC'13-Proceedings of the 2013 ACM Symposium on Theory of Computing*, 665-674, ACM, New York

[25] A. Juditsky and S. Lambert-Lacroix. (2004). Nonparametric confidence set estimation. *Math. Methods Statist.*, **12**(4):410–428.

[26] R. H. Keshavan, A. Montanari, and S. Oh. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.*, **11**:2057–2078.

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

[28] O. Klopp. (2015). Matrix completion by singular value thresholding: sharp bounds. *Electron. J. Statist.*, **9**(2):2348–2369.

[29] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, **39**(5):2302–2329.

[30] M. G. Low. (1997). On nonparametric confidence intervals. *Ann. Statist.*, **25**(6):2547–2554.

[31] S. Negahban and M. J. Wainwright. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.*, **13**:1665–1697.

[32] R. Nickl and B. Szabó (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Process. Appl.* **126**(12), 3913-3934.

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

[34] B. Recht. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.*, **12**:3413–3430.

[35] J. Robins and A. W. van der Vaart. (2007). Adaptive nonparametric confidence sets. *Ann. Statist.*, **34**(1):229–253.

[36] A. Rohde and A. Tsybakov. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, **39**(2):887–930.

[37] A. Singer. (2008). A remark on global positioning from local distances. *Proc. Natl. Acad. Sci. U.S.A.*, **105**(28):9507–9511.

[38] B. Szabó, A. van der Vaart and H. van Zanten (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43**(4):1391–1428.

[39] M. Talagrand. (1996). New concentration inequalities in product spaces. *Invent. Math.*, **126**(3):505–563.