Supplementary Material to "Optimal Link Prediction with Matrix Logistic Regression"

Nicolai Baldin
University of Cambridge

Quentin Berthet
Google Research, Brain Team

A 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

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

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

$$2 \log_2 n \leq k \leq c\sqrt{n}\,. \tag{A.2}$$

The conjecture that no polynomial-time algorithm exists for distinguishing between hypotheses in (A.1) in the regime (A.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, 2013; Wang et al., 2016b; Gao et al., 2017; Cai and Wu, 2018).

The Planted Clique problem can be reduced to the so-called dense subgraph detection problem of testing the null hypothesis in (A.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 1 (The dense subgraph detection conjecture)**.** *For any sequence $k = k_n$ such that $k \leq n^\beta$ 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*

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

The formulation of this conjecture is taken from (Wang et al., 2016a), see their Assumption (A1).

B Proofs

B.1 Some geometric properties of the likelihood

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

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

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 $\sigma^2 = 1/4$,

$$\begin{aligned}\varphi_{\langle\!\langle \zeta, B\rangle\!\rangle_F}(t) &:= \mathbb{E}\big[\exp(t\langle\!\langle \zeta, B\rangle\!\rangle_F)\big] = \prod_{(i,j)\in\Omega} \mathbb{E}\big[\exp(t\varepsilon_{i,j}\langle\!\langle X_j X_i^\top, B\rangle\!\rangle_F)\big] \\ &\le \prod_{(i,j)\in\Omega} \exp\big(t^2\sigma^2 \langle\!\langle X_j X_i^\top, B\rangle\!\rangle_F^{\,2}/2\big) = \exp\big(t\sigma^2 \|\mathbb{X}^\top B \mathbb{X}\|_{F,\Omega}^2/2\big)\,. \end{aligned} \tag{B.1}$$

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 (2.5), and holds for any $\Theta \in \mathcal{P}(M)$,

$$\begin{aligned}\ell(\Theta_\star) - \ell(\Theta) &= \frac{1}{2}\sum_{(i,j)\in\Omega} \big(\sigma'(X_i^\top \Theta_0 X_j)\langle\!\langle X_j X_i^\top, \Theta_\star - \Theta\rangle\!\rangle^2\big) \\ &\ge \frac{\mathcal{L}}{2}\sum_{(i,j)\in\Omega} \langle\!\langle X_j X_i^\top, \Theta_\star - \Theta\rangle\!\rangle_F^2 = \frac{\mathcal{L}}{2}\|\mathbb{X}^\top(\Theta_\star - \Theta)\mathbb{X}\|_{F,\Omega}^2\,, \end{aligned} \tag{B.2}$$

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

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

We shall also be using the bounds

$$\max_{(i,j)\in\Omega} \big(\varepsilon_{i,j} X_i^T(\Theta - \Theta_\star)X_j\big) \le \|\mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\|_{F,\Omega}\,, \quad \text{a.s.}\,, \tag{B.3}$$

$$\mathrm{Var}\big(\langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\rangle\!\rangle_F\big) \le \frac{1}{4}\|\mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X}\|_{F,\Omega}^2\,. \tag{B.4}$$

B.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 *$\varepsilon$-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 2. *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 $\varepsilon > 0$*

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

B.3 Proof of Theorem 10 and Theorem 15

It suffices to show the following uniform deviation inequality

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

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 (3.3) of Theorem 10 the follows directly integrating the deviation inequality (B.5), while the upper bound on

the prediction error in Theorem 15 further follows using (B.2) and the smoothness of the logistic function, $\sup_{t\in\mathbf{R}}\sigma'(t)\le 1/4$. Define

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

The inequality (B.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,

$$\begin{aligned}&\mathbf{P}_{\Theta_\star}\Big(\sup_{\tau(\Theta;\Theta_\star)\ge R_t}\frac{\langle\!\langle\mathcal{E}_\Omega,\mathbb{X}^\top(\Theta-\Theta_\star)\mathbb{X}\rangle\!\rangle}{\tau^2(\Theta;\Theta_\star)}\ge\frac{1}{2}\Big)\\&\quad\le\sum_{K=1}^{d}\sum_{R=1}^{K}\sum_{s=1}^{\infty}\mathbf{P}_{\Theta_\star}\Big(\sup_{\substack{\Theta\in G_{2^sR_t}(\Theta_\star)\\k(\Theta)=K,\,\mathbf{rank}(\Theta)=R}}\langle\!\langle\mathcal{E}_\Omega,\mathbb{X}^\top(\Theta-\Theta_\star)\mathbb{X}\rangle\!\rangle\ge\frac{1}{8}2^{2s}R_t^2\Big)\,.\end{aligned} \tag{B.7}$$

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

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

and therefore using (B.2), this implies

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

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 2 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

$$\begin{aligned}&\mathbb{E}\big[\sup_{\substack{\Theta\in G_{2^sR_t}(\Theta_\star)\\k(\Theta)=K,\,\mathbf{rank}(\Theta)=R}}\langle\!\langle\mathcal{E}_\Omega,\mathbb{X}^\top(\Theta-\Theta_\star)\mathbb{X}\rangle\!\rangle\big]\le C'\int_0^{Z(K,R,s)}\sqrt{H(\varepsilon,\mathbb{T},\|\cdot\|_\Omega)}\mathrm{d}\varepsilon\\&\le C''\sqrt{kr}\int_0^{9Z(K,R,s)}\sqrt{\log\Big(\frac{9Z(K,R,s)}{\varepsilon}\Big)}\mathrm{d}\varepsilon+9C''Z(K,R,s)\sqrt{K\log\big(\frac{de}{K}\big)}\\&\le CZ(K,R,s)t(K,R)\,,\end{aligned}$$

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

$$\mathbf{P}_{\Theta_\star}\Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta_\star) \\ k(\Theta)=K,\, \mathbf{rank}(\Theta)=R}} \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle \geq CZ(K,R,s)t(K,R) \\ + \sqrt{\Big(\frac{1}{2}Z^2(K,R,s) + 4CZ^2(K,R,s)t(K,R)\Big)u} + \frac{Z(K,R,s)u}{3}\Big) \leq e^{-u}\,.$$

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 $\sqrt{c_1+c_2} \leq \sqrt{c_1} + \sqrt{c_2}$ and $\sqrt{c_1 c_2} \leq \frac{1}{2}(c_1\varepsilon + \frac{c_2}{\varepsilon})$, which hold for any $c_1, c_2, \varepsilon > 0$, we obtain

$$\mathbf{P}_{\Theta_\star}\Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta_\star) \\ k(\Theta)=K,\, \mathbf{rank}(\Theta)=R}} \langle\!\langle \mathcal{E}_\Omega, \mathbb{X}^\top(\Theta - \Theta_\star)\mathbb{X} \rangle\!\rangle \\ \geq \frac{1}{16}\mathcal{L}Z^2(K,R,s) + C_1^2 t^2(K,R)/\mathcal{L}\Big) \leq e^{-u(K,R,s)}\,,$$

for some numeric constant $C_1 > 0$. Plugging this back into (B.7) and using (B.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

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

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 (B.9), which concludes the proof.

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

Theorem 3 (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

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

B.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 (B.2), we derive

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

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

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

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

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

B.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 4 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 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 $\mathbf{P}_{\Theta_v}$, $\Theta_u,\Theta_v\in\mathcal{B}^\circ$, $u\ne v$, is upper bounded as

$$\begin{aligned}\mathsf{KL}(\mathbf{P}_{\Theta_u},\mathbf{P}_{\Theta_v})=\mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_u)]-\mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_v)]&\le\frac{1}{8}\sum_{(i,j)\in\Omega}\langle\!\langle X_jX_i^\top,\Theta_u-\Theta_v\rangle\!\rangle_F^2\\&\le\frac{1+\Delta_{\Omega,2k}(\mathbb{X})}{8}k^2\alpha_N^2N\,.\end{aligned}$$

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

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

which, in view of Proposition 4, 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 4.1 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 $\Theta_u,\Theta_v\in\mathcal{G}_k^{\alpha_N}$, $u\ne v$. Furthermore, according to the variant of the Varshamov-Gilbert lemma given as Lemma 5, there exists a subset $\mathcal{G}_k^{\alpha_N,0}\subset\mathcal{G}_k^{\alpha_N}$ such that

$$c_0k^2\alpha_N^2\le\|\Theta_u-\Theta_v\|_F^2\le 2k^2\alpha_N^2\,,$$

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

$$\begin{aligned}\mathsf{KL}(\mathbf{P}_{\Theta_u}, \mathbf{P}_{\Theta_v}) = \mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_u)] - \mathbb{E}_{\mathbf{P}_{\Theta_u}}[\ell_Y(\Theta_v)] &\leq \frac{1}{8}\sum_{(i,j)\in\Omega} \langle\!\langle X_j X_i^\top, \Theta_u - \Theta_v \rangle\!\rangle_F^2 \\ &\leq \frac{1 + \Delta_{\Omega,2k}(\mathbb{X})}{4} k^2 \alpha_N^2 N\,,\end{aligned}$$

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

$$\begin{aligned}\frac{1 + \Delta_{\Omega,2k}(\mathbb{X})}{4} k^2 \alpha_N^2 N + \log 2 = k\gamma \log(2) \log\big(\frac{de}{k}\big) + \log 2 &= \log(2^{k\gamma \log(de/k)+1}) \\ &< \log(2^{\rho k \log(de/k)} + 1)\,.\end{aligned}$$

The desired lower bound then follows from Proposition 4. □

Proposition 4 (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 5 (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*

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

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 of the Variant of the Varshamov-Gilbert lemma. 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

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

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 (B.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 (B.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

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

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 (B.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). □

Proof of Theorem 20. We here provide a proof of the computational lower bound on the prediction error (4.3) for convenience. The bound on the estimation error (4.2) 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

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

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

$$\frac{1}{N}\|\mathbb{X}^\top(\hat{\Theta} - \Theta_\star)\mathbb{X}\|_{F,\Omega}^2 \leq u f(k, d, N), \tag{B.13}$$

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

$$\frac{1}{N}\|\mathbb{X}^\top(\hat{\Theta} - \Theta_\star)\mathbb{X}\|_{F,\Omega}^2 = \sum_{(i,j)\in\Omega} (\hat{\Theta}_{ij} - \Theta_{\star ij})^2 = \|\hat{\Theta} - \Theta_\star\|_{F,\Omega}^2, \tag{B.14}$$

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

$$H_0 : Y \sim \mathbf{P}_0 \quad \text{vs.} \quad H_1 : Y \sim \mathbf{P}_\Theta, \Theta \in \mathcal{G}_k^{\alpha_N}, \tag{B.15}$$

it is natural to employ the following test

$$\psi(Y) = \mathbf{1}\big(\|\hat{\Theta}\|_{F,\Omega} \geq \tau_{d,k}(u)\big), \tag{B.16}$$

where $\tau_{d,k}^2(u) = u f(k, d, N)$. The type I error of this test is controlled automatically due to (B.13) and (B.14), $\mathbf{P}_0(\psi = 1) \leq b/u$. For the type II error, we obtain

$$\begin{aligned}\sup_{\Theta\in\mathcal{G}_k^{\alpha_N}} \mathbf{P}_\Theta(\psi = 0) &= \sup_{\Theta\in\mathcal{G}_k^{\alpha_N}} \mathbf{P}_\Theta\big(\|\hat{\Theta}\|_{F,\Omega} < \tau_{d,k}(u)\big) \\ &\leq \sup_{\Theta\in\mathcal{G}_k^{\alpha_N}} \mathbf{P}_\Theta\big(\|\hat{\Theta} - \Theta\|_{F,\Omega}^2 > \|\Theta\|_{F,\Omega}^2 - \tau_{d,k}^2(u)\big) \leq b/u,\end{aligned}$$

provided that

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

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 (B.16) separates the hypotheses (B.15). This contradicts Conjecture 1 and implies (4.3).

□

Acknowledgements

Nicolai Baldin is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 647812. Quentin Berthet is supported by an Isaac Newton Trust Early Career Support Scheme and by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Alon, N., Krivelevich, M., and Sudakov, B. (1998). Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466.

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

Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

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(4):2342–2359.

Dudley, R. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

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.

Grimmett, G. and McDiarmid, C. (1975). On colouring random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 77, pages 313–324. Cambridge Univ Press.

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

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

Massart, P. (2007). *Concentration inequalities and model selection*, volume 6. Springer.

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

Wang, T., Berthet, Q., and Plan, Y. (2016a). Average-case hardness of rip certification. *arXiv preprint arXiv:1605.09646*.

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