Average-case hardness of RIP certification

Tengyao Wang
Centre for Mathematical Sciences
Cambridge, CB3 0WB, United Kingdom
t.wang@statslab.cam.ac.uk

Quentin Berthet
Centre for Mathematical Sciences
Cambridge, CB3 0WB, United Kingdom
q.berthet@statslab.cam.ac.uk

Yaniv Plan
1986 Mathematics Road
Vancouver BC V6T 1Z2, Canada
yaniv@math.ubc.ca

Abstract

The restricted isometry property (RIP) for design matrices gives guarantees for optimal recovery in sparse linear models. It is of high interest in compressed sensing and statistical learning. This property is particularly important for computationally efficient recovery methods. As a consequence, even though it is in general NP-hard to check that RIP holds, there have been substantial efforts to find tractable proxies for it. These would allow the construction of RIP matrices and the polynomial-time verification of RIP given an arbitrary matrix. We consider the framework of average-case certifiers, that never wrongly declare that a matrix is RIP, while being often correct for random instances. While there are such functions which are tractable in a suboptimal parameter regime, we show that this is a computationally hard task in any better regime. Our results are based on a new, weaker assumption on the problem of detecting dense subgraphs.

Introduction

In many areas of data science, high-dimensional signals contain rich structure. It is of great interest to leverage this structure to improve our ability to describe characteristics of the signal and to make future predictions. Sparsity is a structure of wide applicability (see, e.g. Mallat, 1999; Rauhut and Foucart, 2013; Eldar and Kutyniok, 2012), with a broad literature dedicated to its study in various scientific fields.

The sparse linear model takes the form $y = X\beta + \varepsilon$, where $y \in \mathbf{R}^n$ is a vector of observations, $X \in \mathbf{R}^{n\times p}$ is a *design matrix*, $\varepsilon \in \mathbf{R}^n$ is noise, and the vector $\beta \in \mathbf{R}^p$ is assumed to have a small number k of non-zero entries. Estimating β or the *mean response*, $X\beta$, are among the most widely studied problems in signal processing, as well as in statistical learning. In high-dimensional problems, one would wish to recover β with as few observations as possible. For an incoherent design matrix, it is known that an order of k^2 observations suffice (Donoho, Elad and Temlyakov, 2006; Donoho and Elad, 2003). However, this appears to require a number of observations far exceeding the information content of β, which has only k variables, albeit with unknown locations.

This dependence in k can be greatly improved by using design matrices that are almost isometries on some low dimensional subspaces, i.e., matrices that satisfy the *restricted isometry property* with parameters k and θ, or RIP(k, θ) (see Definition 1.1). It is a highly robust property, and in fact implies that many different polynomial time methods, such as greedy methods (Blumensath and Davies, 2009; Needell and Tropp, 2009; Dai and Milenkovic, 2009) and convex optimization (Candès, 2008; Candès, Romberg and Tao, 2006b; Candès and Tao, 2005), are stable in recovering β.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Random matrices are known to satisfy the RIP when the number n of observation is more than about $k\log(p)/\theta^2$. These results were developed in the field of *compressed sensing* (Candès, Romberg and Tao, 2006a; Donoho, 2006; Rauhut and Foucart, 2013; Eldar and Kutyniok, 2012) where the use of randomness still remains pivotal for near-optimal results. Properties related to the conditioning of design matrices have also been shown to play a key role in the statistical properties of computationally efficient estimators of β (Zhang, Wainwright and Jordan, 2014). While the assumption of randomness allows great theoretical leaps, it leaves open questions for practitioners.

Scientists working on data closely following this model cannot always choose their design matrix X, or at least choose one that is completely random. Moreover, it is in general practically impossible to check that a given matrix satisfies these desired properties, as RIP certification is NP-hard (Bandeira et al., 2012). Having access to a function, or statistic, of X that could be easily computed, which determines how well β may be estimated, would therefore be of a great help. The search for such statistics has been of great importance for over a decade now, and several have been proposed (d'Aspremont and El Ghaoui, 2011; Lee and Bresler, 2008; Juditsky and Nemirovski, 2011; d'Aspremont, Bach and El Ghaoui, 2008). Perhaps the simplest and most popular is the *incoherence parameter*, which measures the maximum inner product between distinct, normalized, columns of X. However, all of these are known to necessarily fail to guarantee good recovery when $p \geq 2n$ unless n is of order k^2 (d'Aspremont and El Ghaoui, 2011). Given a specific problem instance, the strong recovery guarantees of compressed sensing cannot be verified based on these statistics.

In this article, we study the problem of *average-case certification* of the Restricted Isometry Property (RIP). A certifier takes as input a design matrix X, always outputs 'false' when X does not satisfy the property, and outputs 'true' for a large proportion of matrices (see Definition 2.1). Indeed, worst-case hardness does not preclude a problem from being solvable for most instances. The link between restricted isometry and incoherence implies that polynomial time certifiers exists in a regime where n is of order $k^2\log(p)/\theta^2$. It is natural to ask whether the RIP can be certified for sample size $n \gg k\log(p)/\theta^2$, where most matrices (with respect to, say, the Gaussian measure) are RIP. If it does, it would also provide a Las Vegas algorithm to construct RIP design matrices of optimal sizes. This should be compared with the currently existing limitations for the deterministic construction of RIP matrices.

Our main result is that certification in this sense is hard even in a near-optimal regime, assuming a new, weaker assumption on detecting dense subgraphs, related to the *Planted Clique* hypothesis.

Theorem (Informal). *For any $\alpha < 1$, there is no computationally efficient, average-case certifier for the class RIP$_{n,p}(k,\theta)$ uniformly over an asymptotic regime where $n \ll k^{1+\alpha}/\theta^2$.*

This suggests that even in the average case, RIP certification requires almost $k^2\log(p)/\theta^2$ observations. This contrasts highly with the fact that a random matrix satisfies RIP with high probability when n exceeds about $k\log(p)/\theta^2$. Thus, there appears to be a large gap between what a practitioner may be able to certify given a specific problem instance, and what holds for a random matrix.On the other hand, if a certifier is found which fills this gap, the result would not only have huge practical implications in compressed sensing and statistical learning, but would also disprove a long-standing conjecture from computational complexity theory.

We focus solely on the restricted isometry property, but other conditions under which compressed sensing is possible are also known. Extending our results to the restricted eigenvalue condition Bickel, Ritov and Tsybakov (2009) or other conditions (see, van de Geer and Buhlmann, 2009, and references therein) is an interesting path for future research.

Our result shares many characteristics with a hypothesis by Feige (2002) on the hardness of refuting random satisfiability formulas. Indeed, our statement is also about the hardness of verifying that a property holds for a particular instance (RIP for design matrices, instead of unsatisfiability for boolean formulas). It concerns a regime where such a property should hold with high probability (n of order $k^{1+\alpha}/\theta^2$, linear regime for satisfiability), cautiously allowing only one type of errors, false negatives, for a problem that is hard in the worst case. In these two examples, such certifiers exist in a sub-optimal regime. Our problem is conceptually different from results regarding the worst-case hardness of certifying this property (see, e.g. Bandeira et al., 2012; Koiran and Zouzias, 2012; Tillmann and Pfetsch, 2014). It is closer to another line of work concerned with computational lower bounds for statistical learning problems based on average-case assumptions. The planted clique assumption has been used to prove computational hardness results for statistical problems such as estimation and testing of sparse principal components (Berthet and Rigollet, 2013a,b; Wang, Berthet and Samworth, 2016), testing and localization of submatrix signals (Ma and Wu, 2013; Chen and Xu, 2014), community detection (Hajek, Wu and Xu, 2015) and sparse canonical correlation analysis (Gao, Ma and Zhou, 2014). The intractability of noisy parity recovery problem (Blum, Kalai and Wasserman, 2003) has also been used recently as an average-case assumption to deduce computational hardness of detection of satisfiability formulas with lightly planted solutions (Berthet and Ellenberg, 2015). Additionally, several unconditional computational hardness results are shown for statistical problems under constraints of learning models (Feldman et al., 2013). The present work has two main differences compared to previous computational lower bound results. First, in a detection setting, these lower bounds concern two specific distributions (for the null and alternative hypothesis), while ours is valid for all sub-Gaussian distributions, and there is no alternative distribution. Secondly, our result is not based on the usual assumption for the Planted Clique problem. Instead, we use a weaker assumption on a problem of detecting planted dense graphs. This does not mean that the planted graph is a random graph with edge probability $q > 1/2$ as considered in (Arias-Castro and Verzelen, 2013; Bhaskara et al., 2010; Awasthi et al., 2015), but that it can be *any graph* with an unexpectedly high number of edges (see section 3.1). This choice is made to strengthen our result: it would 'survive' the discovery of an algorithm that would use very specific properties of cliques (or even of random dense graphs) to detect their presence. As a consequence, the analysis of our reduction is more technically complicated.

Our work is organized in the following manner: We recall in Section 1 the definition of the restricted isometry property, and some of its known properties. In Section 2, we define the notion of certifier, and prove the existence of a computationally efficient certifier in a sub-optimal regime. Our main result is developed in Section 3, focused on the hardness of average-case certification. The proofs of the main results are in Appendix A of the supplementary material and those of auxiliary results in Appendix B of the the supplementary material.

1 Restricted Isometric Property

1.1 Formulation

We use the definition of Candès and Tao (2005), who introduced this notion. Below, for a vector $u \in \mathbf{R}^p$, we define $\|u\|_0$ is the number of its non-zero entries.

Definition (RIP). A matrix $X \in \mathbb{R}^{n\times p}$ satisfies the *restricted isometry property* with sparsity $k \in \{1, \dots, p\}$ and distortion $\theta \in (0, 1)$, denoted by $X \in \mathrm{RIP}_{n,p}(k, \theta)$, if it holds that

$$1 - \theta \leq \|Xu\|_2^2 \leq 1 + \theta,$$

for every $u \in \mathbb{S}^{p-1}(k) := \{u \in \mathbb{R}^p : \|u\|_2 = 1, \|u\|_0 \leq k\}$.

This can be equivalently defined by a property on submatrices of the design matrix: X is in $\mathrm{RIP}_{n,p}(k, \theta)$ if and only if for any set S of k columns of X, the submatrix, X_{*S}, formed by taking any these columns is almost an isometry, i.e. if the spectrum of its Gram matrix is contained in the interval $[1 - \theta, 1 + \theta]$:

$$\|X_{*S}^\top X_{*S} - I_k\|_{\mathrm{op}} \leq \theta\,.$$

Denote by $\|\cdot\|_{\mathrm{op},k}$ the k*-sparse operator norm*, defined for a matrix A as $\|A\|_{\mathrm{op},k} = \sup_{x\in\mathbb{S}^{p-1}(k)} \|Ax\|_2$. This yields another equivalent formulation of the RIP property: $X \in \mathrm{RIP}_{n,p}(k, \theta)$ if and only if

$$\|X^\top X - I_p\|_{\mathrm{op},k} \leq \theta\,.$$

We assume in the following discussion that the distortion parameter θ is upper-bounded by 1. For $v \in \mathbb{R}^p$ and $T \subseteq \{1, \dots, p\}$, we write v_T for the $\#T$-dimensional vector obtained by restricting v to coordinates indexed by T. Similarly, for an $n \times p$ matrix A and subsets $S \subseteq \{1, \dots, n\}$ and $T \subseteq \{1, \dots, p\}$, we write A_{S*} for the submatrix obtained by restricting A to rows indexed by S, A_{*T} for the submatrix obtained by restricting A to columns indexed by T.

1.2 Generation via random design

Matrices that satisfy the restricted isometry property have many interesting applications in high-dimensional statistics and compressed sensing. However, there is no known way to generate them deterministically in general. It is even NP-hard to check whether a given matrix X belongs to $\mathrm{RIP}_{n,p}(k,\theta)$ (see, e.g Bandeira et al., 2012). Several deterministic constructions of RIP matrices exist for sparsity level $k \lesssim \theta\sqrt{n}$. For example, using equitriangular tight frames and Gershgorin's circle theorem, one can construct RIP matrices with sparsity $k \le \sqrt{n}$ and distortion θ bounded away from 0 (see, e.g. Bandeira et al., 2012). The limitation $k \le \theta\sqrt{n}$ is known as the 'square root bottleneck'. To date, the only constructions that break the 'square root bottleneck' are due to Bourgain et al. (2011) and Bandeira, Mixon and Moreira (2014), both of which give RIP guarantee for k of order $n^{1/2+\epsilon}$ for some small $\epsilon > 0$ and fixed θ (the latter construction is conditional on a number-theoretic conjecture being true).

Interestingly though, it is easy to generate large matrices satisfying the restricted isometry property through random design, and compared to the fixed design matrices mentioned in the previous paragraph, these random design constructions are much less restrictive on the sparsity level, typically allowing k up to the order $n/\log(p)$ (assuming θ is bounded away from zero). They can be constructed easily from any centred sub-Gaussian distribution. We recall that a distribution Q (and its associated random variable) is said to be sub-Gaussian with parameter σ if $\int_{\mathbb{R}} e^{\lambda x}\, dQ(x) \le e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$.

Definition. Define $\mathcal{Q} = \mathcal{Q}_\sigma$ to be the set of sub-Gaussian distributions Q over $\mathbb{R}$ with zero mean, unit variance, and sub-Gaussian parameter at most σ.

The most common choice for a $Q \in \mathcal{Q}$ is the standard normal distribution $\mathcal{N}(0,1)$. Note that by Taylor expansion, for any $Q \in \mathcal{Q}$, we necessarily have $\sigma^2 \ge \int_{\mathbb{R}} x^2\, dQ(x) = 1$. In the rest of the paper, we treat σ as fixed. Define the normalized distribution $\tilde{Q}$ to be the distribution of $Z/\sqrt{n}$ for $Z \sim Q$. The following well-known result states that by concentration of measure, random matrices generated with distribution $\tilde{Q}^{\otimes(n\times p)}$ satisfy restricted isometries (see, e.g. Candès and Tao (2005) and Baraniuk et al. (2008)). For completeness, we include a proof that establishes these particular constants stated here. All proofs are deferred to Appendix A or Appendix B of the supplementary material.

Proposition 1. *Suppose X is a random matrix with distribution $\tilde{Q}^{\otimes(n\times p)}$, where $Q \in \mathcal{Q}$. It holds that*

$$\mathbb{P}\big(X \in \mathrm{RIP}_{n,p}(k,\theta)\big) \ge 1 - 2\exp\biggl\{k\log\Bigl(\frac{9ep}{k}\Bigr) - \frac{n\theta^2}{256\sigma^4}\biggr\}. \tag{1}$$

In order to clarify the notion of asymptotic regimes used in this paper, we introduce the following definition.

Definition. For $0 \le \alpha \le 1$, define the asymptotic regime

$$\mathcal{R}_\alpha := \biggl\{(p_n, k_n, \theta_n)_n : p, k \to \infty \text{ and } n \gg \frac{k_n^{1+\alpha}\log(p_n)}{\theta_n^2}\biggr\}.$$

We note that in this notation, Proposition 1 implies that for $(p,k,\theta) = (p_n,k_n,\theta_n) \in \mathcal{R}_0$ we have, $\lim_{n\to\infty} \tilde{Q}^{\otimes(n\times p)}(X \in \mathrm{RIP}_{n,p}(k,\theta)) = 1$, and this convergence is uniform over $Q \in \mathcal{Q}$.

2 Certification of Restricted Isometry

2.1 Objectives and definition

In practice, it is useful to know with certainty whether a particular realization of a random design matrix satisfies the RIP condition. It is known that the problem of deciding if a given matrix is RIP is NP-hard (Bandeira et al., 2012). However, NP-hardness is a only a statement about worst-case instances. It would still be of great use to have an algorithm that can correctly decide RIP property for an average instance of a design matrix, with some accuracy. Such an algorithm should identify a high proportion of RIP matrices generated through random design and make no false positive claims. We call such an algorithm an *average-case certifier*, or a *certifier* for short.

Definition (Certifier)**.** Given a parameter sequence $(p,k,\theta) = (p_n,k_n,\theta_n)$, we define a *certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices* to be a sequence $(\psi_n)_n$ of measurable functions $\psi_n : \mathbb{R}^{n\times p} \to \{0,1\}$, such that

$$\psi_n^{-1}(1) \subseteq \mathrm{RIP}_{n,p}(k,\theta) \qquad \text{and} \qquad \limsup_{n\to\infty} \tilde{Q}^{\otimes(n\times p)}\big(\psi_n^{-1}(0)\big) \le 1/3. \tag{2}$$

Note the definition of a certifier depends on both the asymptotic parameter sequence (p_n, k_n, θ_n) and the sub-Gaussian distribution Q. However, when it is clear from the context, we will suppress the dependence and refer to certifiers for $\mathrm{RIP}_{n,p}(k,\theta)$ properties of $\tilde{Q}^{\otimes(n\times p)}$-random matrices simply as 'certifiers'.

The two defining properties in (2) can be understood as follows. The first condition means that if a certifier outputs 1, we know with certainty that the matrix is RIP. The second condition means that the certifier is not overly conservative; it is allowed to output 0 for at most one third (with respect to $\tilde{Q}^{\otimes(n\times p)}$ measure) of the matrices. The choice of $1/3$ in the definition of a certifier is made to simplify proofs. However, all subsequent results will still hold if we replace $1/3$ by any constant in $(0,1)$. In view of Proposition 1, the second condition in (2) can be equivalently stated as

$$\liminf_{n\to\infty} \tilde{Q}^{\otimes(n\times p)}\big\{\psi_n(X) = 1 \bigm| X \in \mathrm{RIP}_{n,p}(k,\theta)\big\} \geq 2/3.$$

With such a certifier, given an arbitrary problem fitting the sparse linear model, the matrix X could be tested for the restricted isometry property, with some expectation of a positive result. This would be particularly interesting given a certifier in the parameter regime $n \ll \theta_n^2 k_n^2$, in which presently known polynomial-time certifiers cannot give positive results.

Even though it is not the main focus of our paper, we also note that a certifier ψ with the above properties for some distribution $Q \in \mathcal{Q}$ would form a certifier/distribution couple (ψ, Q), that yields in the usual manner a Las Vegas algorithm to generate RIP matrices. The (random) algorithm keeps generating random matrices $X \sim \tilde{Q}^{\otimes(n\times p)}$ until $\psi_n(X) = 1$. The number of times that the certifier is invoked has a geometric distribution with success probability $\tilde{Q}^{\otimes(n\times p)}\big(\psi_n^{-1}(1)\big)$. Hence, the Las Vegas algorithm runs in randomized polynomial time if and only if ψ_n runs in randomized polynomial time.

2.2 Certifier properties

Although our focus is on algorithmically efficient certifiers, we establish first the properties of a certifier that is computationally intractable. This certifier serves as a benchmark for the performance of other candidates. Indeed, we exhibit in the following proposition a certifier, based on the k-sparse operator norm, that works uniformly well in the same asymptotic parameter regime $\mathcal{R}_0$, where $\tilde{Q}^{\otimes(n\times p)}$-random matrices are RIP with asymptotic probability 1. For clarity, we stress that our criterion when judging a certifier will always be its uniform performance over asymptotic regimes $\mathcal{R}_\alpha$ for some $\alpha \in [0,1]$.

Proposition 2. *Suppose $(p,k,\theta) = (p_n,k_n,\theta_n) \in \mathcal{R}_0$. Furthermore, Let $Q \in \mathcal{Q}$ and $X \sim \tilde{Q}^{\otimes(n\times p)}$. Then the sequence of tests $(\psi_{op,k})_n$ based on sparse operator norms, defined by*

$$\psi_{op,k}(X) := \mathbf{1}\Big\{\|X^\top X - I_p\|_{\mathrm{op},k} \leq \theta\Big\}.$$

is a certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices.

By a direct reduction from the clique problem, one can show that it is NP-hard to compute the k-sparse operator norm of a matrix. Hence the certifier $\psi_{\mathrm{op},k}$ is computationally intractable. The next proposition concerns the certifier property of a test based on the maximum incoherence between columns of the design matrix. It follows directly from a well-known result on the incoherence parameter of a random matrix (see, e.g. Rauhut and Foucart (2013, Proposition 6.2)) and allows the construction of a polynomial-time certifier that works uniformly well in the asymptotic parameter regime $\mathcal{R}_1$.

Proposition 3. *Suppose $(p,k,\theta) = (p_n,k_n,\theta_n)$ satisfies $n \geq 196\sigma^4 k^2 \log(p)/\theta^2$. Let $Q \in \mathcal{Q}$ and $X \sim \tilde{Q}^{\otimes(n\times p)}$, then the tests ψ_∞ defined by*

$$\psi_\infty(X) := \mathbf{1}\left\{\|X^\top X - I_p\|_\infty \leq 14\sigma^2\sqrt{\frac{\log(p)}{n}}\right\}$$

is a certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices.

Proposition 3 shows that, when the sample size n is above $k^2\log(p)/\theta^2$ in magnitude (in particular, this is satisfied asymptotically when $(p,k,\theta)=(p_n,k_n,\theta_n)\in\mathcal{R}_1$), there is a polynomial time certifier. In other words, in this high-signal regime, the average-case decision problem for RIP property is much more tractable than indicated by the worst-case result. On the other hand, the certifier in Proposition 3 works in a much smaller parameter range when compared to $\psi_{\mathrm{op},k}$ in Proposition 2. Combining Proposition 2 and 3, we have the following schematic diagram (Figure 1). When the sample size is lower than specified in $\mathcal{R}_0$, the property does not hold, with high probability, and no certifier exists. A computationally intractable certifier works uniformly over $\mathcal{R}_0$. On the other end of the spectrum, when the sample size is large enough to be in $\mathcal{R}_1$, a simple certifier based on the maximum incoherence of the design matrix is known to work in polynomial time. This leaves open the question of whether (randomized) polynomial time certifiers can work uniformly well in $\mathcal{R}_0$, or $\mathcal{R}_\alpha$ for any $\alpha\in[0,1)$. We will see in the next section that, assuming a weaker variant of the Planted Clique hypothesis from computational complexity theory, $\mathcal{R}_1$ is essentially the largest asymptotic regime where a randomized polynomial time certifier can exist.

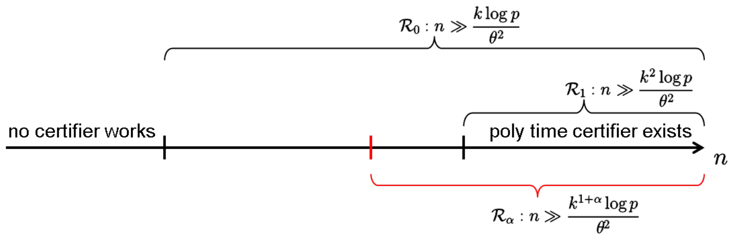
$\mathcal{R}_0 : n \gg \frac{k\log p}{\theta^2}$
$\mathcal{R}_1 : n \gg \frac{k^2\log p}{\theta^2}$
no certifier works
poly time certifier exists
n
$\mathcal{R}_\alpha : n \gg \frac{k^{1+\alpha}\log p}{\theta^2}$

Figure 1: Schematic digram for existence of certifiers in different asymptotic regimes.

3 Hardness of Certification

3.1 Planted dense subgraph assumptions

We show in this section that certification of RIP property is an average-case hard problem in the parameter regime $\mathcal{R}_\alpha$ for any $\alpha<1$. This is precisely the regime not covered by Proposition 3. The average-case hardness result is proved via reduction to the planted dense subgraph assumption.

For any integer $m\geq 0$, denote $\mathbb{G}_m$ the collection of all graphs on m vertices. We write $V(G)$ and $E(G)$ for the set of vertices and edges of a graph G. For $H\in\mathbb{G}_\kappa$ where $\kappa\in\{0,\ldots,m\}$, let $\mathcal{G}(m,1/2,H)$ be the random graph model that generates a random graph G on m vertices as follows. It first picks κ random vertices $K\subseteq V(G)$ and plants an isomorphic copy of H on these κ vertices, then every pair of vertices not in $K\times K$ is connected by an edge independently with probability $1/2$. We write $\mathbf{P}_H$ for the probability measure on $\mathbb{G}_m$ associated with $\mathcal{G}(m,1/2,H)$. Note that if H is the empty graph, then $\mathcal{G}(m,1/2,\emptyset)$ describes the Erdős–Rényi random graph. With a slight abuse of notation, we write $\mathbf{P}_0$ in place of $\mathbf{P}_\emptyset$. On the other hand, for $\epsilon\in(0,1/2]$, if H belongs to the set

$$\mathcal{H}=\mathcal{H}_{\kappa,\epsilon}:=\left\{H\in\mathbb{G}_\kappa:\#E(H)\geq(1/2+\epsilon)\frac{\kappa(\kappa-1)}{2}\right\},$$

then $\mathcal{G}(m,1/2,H)$ generates random graphs that contain elevated local edge density. The planted dense graph problem concerns testing apart the following two hypotheses:

$$H_0: G\sim\mathcal{G}(m,1/2,\emptyset)\qquad\text{and}\qquad H_1: G\sim\mathcal{G}(m,1/2,H)\text{ for some }H\in\mathcal{H}_{\kappa,\epsilon}. \tag{3}$$

It is widely believed that for $\kappa=O(m^{1/2-\delta})$, there does not exist randomized polynomial time tests to distinguish between H_0 and H_1 (see, e.g. Jerrum (1992); Feige and Krauthgamer (2003); Feldman et al. (2013)). More precisely, we have the following assumption.

Assumption (A1) 1. Fix $\epsilon\in(0,1/2]$ and $\delta\in(0,1/2)$. let $(\kappa_m)_m$ be any sequence of integers such that $\kappa_m\to\infty$ and $\kappa_m=O\big(m^{1/2-\delta}\big)$. For any sequence of randomized polynomial time tests $(\phi_m:\mathbb{G}_m\to\{0,1\})_m$, we have

$$\liminf_m\Big\{\mathbf{P}_0\big(\phi(G)=1\big)+\max_{H\in\mathcal{H}_{\kappa,\epsilon}}\mathbf{P}_H\big(\phi(G)=0\big)\Big\}>1/3\,.$$

We remark that if $\epsilon = 1/2$, then $\mathcal{H}_{\kappa,\epsilon}$ contains only the κ-complete graph and the testing problem becomes the well-known planted clique problem (cf. Jerrum (1992) and references in Berthet and Rigollet (2013a,b)).

The difficulty of this problem has been used as a primitive for the hardness of other tasks, such as cryptographic applications, in Juels and Peinado (2000), testing for k-wise dependence in Alon et al. (2007), approximating Nash equilibria in Hazan and Krauthgamer (2011). In this case, Assumption **(A1)** is a version of the planted clique hypothesis (see, e.g. Berthet and Rigollet (2013b, Assumption $\mathbf{A}_{\mathrm{PC}}$)). We emphasize that Assumption A1 is significantly milder than the planted clique hypothesis (since it allows any $\epsilon \in (0, 1/2]$), or that a hypothesis on planted random graphs. We also note that when $\kappa \geq C_\epsilon \sqrt{m}$, spectral methods can be used to detect such graphs with high probability. Indeed, when G contains a graph of $\mathcal{H}$, denoting A_G its adjacency matrix, then $A_G - \mathbf{1}\mathbf{1}^\top/2$ has a leading eigenvalue greater than $\epsilon(\kappa - 1)$, whereas it is of order $\sqrt{m}$ for a usual Erdős–Rényi random graph.

The following theorem relates the hardness of the planted dense subgraph testing problem to the hardness of certifying restricted isometry of random matrices. We recall that the distribution of X is that of an $n \times p$ random matrix with entries independently and identically sampled from $\tilde{Q} \overset{d}{=} Q/\sqrt{n}$, for some $Q \in \mathcal{Q}$. We also write Ψ_{rp} for the class of randomized polynomial time certifiers.

Theorem 4. *Assume (A1) and fix any $\alpha \in [0, 1)$. Then there exists a sequence $(p, k, \theta) = (p_n, k_n, \theta_n) \in \mathcal{R}_\alpha$, such that there is no certifier/distribution couple $(\psi, Q) \in \Psi_{\mathrm{rp}} \times \mathcal{Q}$ with respect to this sequence of parameters.*

Our proof of Theorem 4 relies on the following ideas: Given a graph G, an instance of the planted clique problem in the assumed hard regime, we construct n random vectors based on the adjacency matrix of a bipartite subgraph of G, between two random sets of vertices. Each coefficient of these vectors is then randomly drawn from one of two carefully chosen distributions, conditionally on the presence or absence of a particular edge. This construction ensures that if the graph is an Erdős–Rényi random graph (i.e. with no planted graph), the vectors are independent with independent coefficients, with distribution $\tilde{Q}$. Otherwise, we show that with high probability, the presence of an unusually dense subgraph will make it very likely that the matrix does not satisfy the restricted isometry property, for a set of parameters in $\mathcal{R}_\alpha$. As a consequence, if there existed a certifier/distribution couple $(\psi, Q) \in \Psi_{\mathrm{rp}} \times \mathcal{Q}$ in this range of parameters, it could be used - by using as input in the certifier the newly constructed matrix - to determine with high probability the distribution of G, violating our assumption **(A1)**.

We remark that this result holds for *any* distribution in $\mathcal{Q}$, in contrast to computational lower bounds in statistical learning problems, that apply to a specific distribution. For the sake of simplicity, we have kept the coefficients of X identically distributed, but our analysis is not dependent on that fact, and our result can be directly extended to the case where the coefficients are independent, with different distributions in $\mathcal{Q}$.

Theorem 4 may be viewed as providing an asymptotic lower bound of the sample size n for the existence of a computationally feasible certifier. It establishes this computational lower bound by exhibiting some specific 'hard' sequences of parameters inside $\mathcal{R}_\alpha$, and show that any algorithm violating the computational lower bound could be exploited to solve the planted dense subgraph problem. All hardness results, whether in a worst-case (NP-hardness, or other) or the average-case (by reduction from a hard problem), are by nature statements on the impossibility of accomplishing a task in a computationally efficient manner, uniformly over a range of parameters. They are therefore always based on the construction of a 'hard' sequence of parameters used in the reduction, for which a contradiction is shown. Here, the 'hard' sequence is explicitly constructed in the proof to be some $(p, k, \theta) = (p_n, k_n, \theta_n)$ satisfying $p \geq n$ and $n^{1/(3-\alpha-4\beta)} \ll k \ll n^{1/(2-\beta)-\delta}$, for $\beta \in [0, (1-\alpha)/3)$ and any small $\delta > 0$. The tuning parameter β is to allow additional flexibility in choosing these 'hard' sequences. More precisely, using an averaging trick first seen in Ma and Wu (2013), we are able to show that the existence of such 'hard' sequences is not confined only in the sparsity regime $k \ll n^{1/2}$. We note that in all our 'hard' sequences, θ_n must depend on n. An interesting extension is to see if similar computational lower bounds hold when restricted to a subset of $\mathcal{R}_\alpha$ where θ is constant.

References

Alon, N., Andoni, A., Kaufman, T., Matulef, K., Rubinfeld, R., and Xie, N. (2007) Testing k-wise and almost k-wise independence. *Proceedings of the Thirty-ninth ACM STOC*. 496–505.

Arias-Castro, E., Verzelen, N. (2013) Community Detection in Dense Random Networks. *Ann. Statist.*,**42**, 940-969

Awasthi, P., Charikar, M., Lai, K. A. and Risteki, A. (2015) Label optimal regret bounds for online local learning. *J. Mach. Learn. Res. (COLT)*, **40**.

Bandeira, A. S., Dobriban, E., Mixon, D. G. and Sawin, W. F. (2012) Certifying the restricted isometry property is hard. *IEEE Trans. Information Theory*, **59**, 3448–3450.

Bandeira, A. S., Mixon, D. G. and Moreira, J. (2014) A conditional construction of restricted isometries. *International Mathematics Research Notices*, to appear.

Baraniuk, R., Davenport, M., DeVore, R. and Wakin, M. (2008) A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, **28**, 253–263.

Berthet, Q. and Ellenberg, J. S. (2015) Detection of Planted Solutions for Flat Satisfiability Problems. *Preprint*

Berthet, Q. and Rigollet P. (2013) Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, **41**, 1780–1815.

Berthet, Q. and Rigollet P. (2013) Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res. (COLT)*, **30**, 1046–1066.

Bhaskara, A., Charikar, M., Chlamtac, E., Feige, U. and Vijayaraghavan, A. (2010) Detecting High Log-Densities an $O(n^{1/4})$ Approximation for Densest k-Subgraph. *Proceedings of the forty-second ACM symposium on Theory of computing*, 201–210.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector *Ann. Statist.*, **37**,1705–1732

Blum, A., Kalai, A. and Wasserman, H. (2003) Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, *50*, 506–519.

Blumensath, T. and Davies, M. E. (2009) Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, **27**, 265–274.

Bourgain, J., Dilworth, S., Ford, K. and Konyagin, S. (2011) Explicit constructions of RIP matrices and related problems. *Duke Math. J.*, **159**, 145–185.

Candès, E. J. (2008) The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, **346**, 589–592.

Candès, E. J., Romberg, J. and Tao, T. (2006) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, **52**, 489–509.

Candès, E. J., Romberg, J. K. and Tao, T. (2006) Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, **59**, 2006.

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

Chen, Y. and Xu, J. (2014) Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *preprint*, arXiv:1402.1267.

d'Aspremont, A., Bach, F. and El Ghaoui, L. (2008) Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, **9**, 1269–1294.

d'Aspremont, A. and El Ghaoui, L. (2011) Testing the nullspace property using semidefinite programming. *Mathematical programming*, **127**, 123–144.

Dai, W. and Milenkovic, O. (2009) Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory*, **55**, 2230–2249.

Donoho, D. L. (2006) Compressed sensing. *IEEE Trans. Inform. Theory*, **52**, 1289–1306.

Donoho, D. L., and Elad, M. (2003) mally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, **100**, 2197–2202.

Donoho, D. L., Elad, M. and Temlyakov, V. N. (2006) Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, **52**, 6–18.

Eldar, Y. C. and Kutyniok, G. (2012) *Compressed Sensing: Theory and Applications*. Cambridge University Press, Cambridge.

Feige, U. Relations between average case complexity and approximation complexity. *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, 534–543.

Feige, U. and Krauthgamer, R. (2003) The probable value of the Lovàsz–Schrijver relaxations for a maximum independent set. *SIAM J. Comput.*, **32**, 345–370.

Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S. S. and Xiao, Y. (2013) Statistical Algorithms and a Lower Bound for Detecting Planted Cliques. *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*. 655–664.

Gao, C., Ma, Z. and Zhou, H. H. (2014) Sparse CCA: adaptive estimation and computational barriers. *preprint*, arXiv:1409.8565.

Hajek, B., Wu, Y. and Xu, J.(2015) Computational Lower Bounds for Community Detection on Random Graphs, *Proceedings of The 28th Conference on Learning Theory*, 899–928.

Hazan, E. and Krauthgamer, R. (2011) How hard is it to approximate the best nash equilibrium? *SIAM J. Comput.*, **40**, 79–91.

Jerrum, M. (1992) Large cliques elude the Metropolis process. *Random Struct. Algor.*, **3**, 347–359.

Juditsky, A. and Nemirovski, A. (2011) On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Mathematical programming*, **127**, 57–88.

Juels, A. and Peinado, M. (2000) Hiding cliques for cryptographic security. *Des. Codes Cryptography*. **20**, 269-280.

Koiran, P. and Zouzias, A. (2012) Hidden cliques and the certification of the restricted isometry property. *preprint*, arXiv:1211.0665.

Lee, K. and Bresler, Y. (2008) Computing performance guarantees for compressed sensing. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 5129–5132.

Ma, Z. and Wu, Y. (2013) Computational barriers in minimax submatrix detection. *arXiv preprint*.

Mallat, S. (1999) *A wavelet tour of signal processing*. Academic press, Cambridge, MA.

Needell, D. and Tropp, J. A. (2009) CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, **26**, 301–321.

Rauhut, H. and Foucart, S. (2013) *A Mathematical Introduction to Compressive Sensing*. Birkhäuser.

Tillmann, A. N. and Pfetsch M. E. (2014) The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, **60**, 1248–1259.

van de Geer, S. and Buhlmann, P. (2009) On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.*, **3**, 1360–1392

Wang, T., Berthet, Q. and Samworth, R. J. (2016) Statistical and computational trade-offs in Estimation of Sparse Pincipal Components. *Ann. Statist.*, **45**, 1896–1930

Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014) Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *JMLR: Workshop and Conference Proceedings (COLT)*, **35**, 921–948.

Supplementary material to “Average-case hardness of RIP certification”

Tengyao Wang
Centre for Mathematical Sciences
Cambridge, CB3 0WB, United Kingdom
t.wang@statslab.cam.ac.uk

Quentin Berthet
Centre for Mathematical Sciences
Cambridge, CB3 0WB, United Kingdom
q.berthet@statslab.cam.ac.uk

Yaniv Plan
1986 Mathematics Road
Vancouver BC V6T 1Z2, Canada
yaniv@math.ubc.ca

A Proofs of Main Results

Proof of Theorem 4. We prove by contradiction. Assume the contrary, that $(\psi_n)_n$ is a polynomial time computable certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices.

For $\alpha < 1$ and $0 \leq \beta < \frac{1}{3}(1-\alpha)$, let $(p,k,\theta) = (p_n,k_n,\theta_n) \in \mathcal{R}_\alpha$ be a sequence satisfying $p \geq n$, $\log p = O(\log n)$, $n^{\frac{1}{3-\alpha-4\beta}} \ll k \ll n^{\frac{1}{2-\beta}-\delta}$ for some $\delta > 0$ and $\theta = \sqrt{k^{1+\alpha}\log(p)/n}$. Let $L = 10$ and $\ell = \lfloor k^\beta \rfloor$. Define $m = L\ell n$ and $\kappa = Lk$. We check that

$$\kappa^2 \asymp k^{2-\beta}k^\beta \ll n^{1-\delta}\ell \ll m^{1-\delta'}$$

for some positive δ' that depends on δ only. We remark that the purpose of introducing the extra parameter β in the proof is mainly to show the ubiquity of parameter sequences (p,k,θ) that arrive at a contradiction. In particular, we can use positive β values to construct sequences where $k \gg n^{1/2}$. For a first reading, it suffices to take $\beta = 0$ (i.e. $\ell = 1$), which already constitutes a proof of the theorem. When $\beta > 0$, the proof requires the additional assumption that there exists $\breve{Q}$ such that for $Y_1,\ldots,Y_\ell^2 \overset{\text{i.i.d.}}{\sim} \breve{Q}$, $\ell^{-1}\sum_{i=1}^{\ell^2} Y_i \sim \tilde{Q}$. Note when $\beta = 0$, we can simply take $\breve{Q} = \tilde{Q}$. Let ξ denote the median of $\breve{Q}$. By definition of the median, there exists a unique decomposition of the probability measure $\breve{Q}$ as $\breve{Q} = \frac{1}{2}\breve{Q}^+ + \frac{1}{2}\breve{Q}^-$, where $\breve{Q}^+$ and $\breve{Q}^-$ are probability measures supported on $(-\infty,\xi]$ and $[\xi,\infty)$ respectively.

We prove below that Algorithm 1, which runs in randomized polynomial time, can distinguish between $\mathbf{P}_0$ and $\mathbf{P}_H$ with zero asymptotic error for any choice of $H \in \mathcal{H}_{\kappa,\epsilon}$.

First, assume $G \sim \mathbf{P}_0$. Then matrix A from Step 1 of Algorithm 1 have independent Rademacher entries, which implies that $X \sim \tilde{Q}^{\otimes(n\times p)}$. Therefore, by (2) in Section 2 we must have

$$\liminf \mathbf{P}_0(\phi(G)=1) = \tilde{Q}^{\otimes(n\times p)}(\psi_n^{-1}(0)) < 1/3.$$

Next, assume G is generated with probability measure $\mathbf{P}_H$ for some $H \in \mathcal{H}_{\kappa,\epsilon}$. We claim that

$$\tilde{X} \notin \mathrm{RIP}_{n,n}\Big(k, \frac{ck^2}{n\ell^2}\Big) \tag{1}$$

for some absolute positive constant c. Since

$$\frac{k^2}{n\ell^2} \gg \sqrt{\frac{k^{1+\alpha}}{n}} \gg \theta,$$

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Algorithm 1: Pseudo-code for an algorithm to distinguish between $\mathbf{P}_0$ and $\mathbf{P}_H$.

```
Input: $m \in \mathbb{N}$, $\kappa \in \{1, \ldots, m\}$, $G \in \mathbb{G}_m$, $L \in \mathbb{N}$
begin
    Step 1: Let $N \leftarrow \lfloor m/L \rfloor$, $k \leftarrow \lfloor \kappa/L \rfloor$, $\ell \leftarrow \lfloor k^\beta \rfloor$, $n \leftarrow \lfloor N/\ell \rfloor$, $p \leftarrow p_n$. Draw
        $u_1, \ldots, u_N, w_1, \ldots, w_N$ uniformly at random without replacement from $V(G)$. Form
        $A = (A_{ij}) \in \mathbb{R}^{N \times N}$ where $A_{ij} = 2 \cdot \mathbb{1}_{\{u_i \sim w_j\}} - 1$.
    Step 2: Let $Y^+ = (Y_{ij}^+)$ and $Y^- = (Y_{ij}^-)$ be $N$-by-$N$ random matrices independent from all
        other random variables and from each other, and such that $Y_{ij}^+ \overset{\text{i.i.d.}}{\sim} \breve{Q}^+$ and $Y_{ij}^- \overset{\text{i.i.d.}}{\sim} \breve{Q}^-$.
        Define $Z = (Z_{ij})$ by $Z_{ij} = \mathbb{1}\{A_{ij} = 1\} Y_{ij}^+ + \mathbb{1}\{A_{ij} = -1\} Y_{ij}^-$.
    Step 3: For $0 \le a, b \le \ell - 1$, define $Z^{(a,b)} \in \mathbb{R}^{n \times n}$ by $Z_{i,j}^{(a,b)} = Z_{an+i, bn+j}$. Define
        $\tilde{X} \leftarrow \ell^{-1} \sum_{0 \le a,b < \ell} Z^{(a,b)}$. Finally, let $X \leftarrow \begin{pmatrix} \tilde{X} & \tilde{X}' \end{pmatrix}$ where $\tilde{X}' \in \mathbb{R}^{n \times (p-n)}$ has entries
        independently drawn from distribution $\tilde{Q}$.
    Step 4: Let $\phi(G) \leftarrow 1 - \psi_n(X)$.
end
Output: $\phi(G)$
```

we have that for large n, $\tilde{X} \notin \mathrm{RIP}_{n,n}(k, \theta)$. Hence X is *a fortiori* not an $\mathrm{RIP}_{n,p}(k, \theta)$ matrix. As a result,

$$\liminf_m \max_{H \in \mathcal{H}_{\kappa,\epsilon}} \mathbf{P}_H\big(\phi(G) = 0\big)\big) < 1/3,$$

contradicting Assumption $(\mathbf{A1})$.

It remains to verify the claimed result in (1). Let $K \subseteq V(G)$ be the κ-subset of vertices on which the subgraph H is planted. We write $U = \{u_1, \ldots, u_N\}$ and $W = \{w_1, \ldots, w_N\}$ for the two random subsets of vertices. Let $N_{U,W;K}$ be the random variable counting the number of edges in G with two endpoints in $U \cap K$ and $W \cap K$ respectively. Then

$$\begin{aligned} N_{U,W;K} &= \#\Big\{\{u, w\} \in E(G) : u \in U \cap K, w \in W \cap K\Big\} \\ &= \sum_{u \in K} \sum_{w \in K} \mathbb{1}\{u \in U\} \mathbb{1}\{w \in W\} \mathbb{1}\{u \sim w\}. \end{aligned}$$

Define

$$\Omega_1 := \left\{ N_{U,W;K} \ge \left(\frac{1}{2} + \frac{\epsilon}{4}\right) k^2 \right\} \cap \left\{ \big|\#U \cap K - k\big| \le \frac{\epsilon}{8} k \right\} \cap \left\{ \big|\#W \cap K - k\big| \le \frac{\epsilon}{8} k \right\}.$$

Lemma 1 below shows that Ω_1 has asymptotic probability 1. Note Ω_1 is in the σ-algebra of (U, W). Let $U = U_0$ and $W = W_0$ be any realization satisfying Ω_1. We write $\mathbb{P}^{U_0,W_0}$ and $\mathbb{E}^{U_0,W_0}$ as shorthand for the probability and expectation conditional on $U = U_0$ and $W = W_0$.

For each $j \in \{1, \ldots, n\}$, define $s_j := \sum_{u_i \in U \cap K} A_{i,j}$. Write $k_1 := (1 - \epsilon/8)k$ and $k_2 = (1 + \epsilon/8)k$. Let $S := \{i : u_i \in U \cap K\}$, and let T be a subset of k_1 indices in $\{1, \ldots, n\}$ corresponding to the k_1 largest values of s_j (breaking ties arbitrarily). Note that S and T are functions of U and V. On the event $U = U_0$ and $W = W_0$, both $\#S = \#U \cap K$ and $\#W \cap K$ are bounded in the interval $[k_1, k_2]$, so in particular $k_1 \le \#W \cap K$. We have

$$\sum_{w_j \in W \cap K} s_j = 2N_{U,W;K} - \#(U \cap K) \times \#(W \cap K) \ge \big\{(1 + \epsilon/2) - (1 + \epsilon/8)^2\big\} k^2 \ge \frac{\epsilon}{5} k^2.$$

As elements of T index columns of A corresponding to largest values of s_js, we have that on event $\{U = U_0, W = W_0\}$,

$$\sum_{j \in T} s_j \ge \frac{\#T}{\#W \cap K} \frac{\epsilon}{5} k^2 \ge \frac{\epsilon}{5} \frac{k^2 k_1}{k_2} \ge \frac{\epsilon}{6} k k_1. \tag{2}$$

Define the unit vector $v \in \mathbb{R}^n$ by $v_T = k_1^{-1/2}\mathbf{1}_{k_1}$ and $v_{T^c} = 0$. Note that v is k_1-sparse and hence also k-sparse. Conditional on $U = U_0$ and $W = W_0$, $Z_{ij} = Y_{ij}^+$ if $A_{ij} = 1$ and $Z_{ij} = Y_{ij}^-$ if $A_{ij} = -1$. By definition of $\tilde{Q}^+$ and $\tilde{Q}^-$, and the fact that $\tilde{Q}$ is not a point mass, we have $\mathbb{E}Y_{ij}^+ = -\mathbb{E}Y_{ij}^- = c_1/\sqrt{n}$ for some absolute constant $c_1 > 0$. By (2), the sum $\sum_{i\in S, j\in T} Z_{ij}$ can be bounded below in conditional expectation by

$$\begin{aligned}\mathbb{E}^{U_0,W_0}\sum_{i\in S,j\in T} Z_{ij} &\geq \mathbb{E}^{U_0,W_0}\biggl(\sum_{i\in S,j\in T}(\mathbb{1}\{A_{ij}=1\}Y_{ij}^+ + \mathbb{1}\{A_{ij}=-1\}Y_{ij}^-)\biggr)\\ &= \frac{c_1}{\sqrt{n}}\biggl(\sum_{j\in T} s_j\biggr) \geq \frac{c_1}{\sqrt{n}}\frac{\epsilon}{6}kk_1\,.\end{aligned}$$

By Lemma 3, both $Y_{ij}^+ - \mathbb{E}Y_{ij}^+$ and $Y_{ij}^- - \mathbb{E}Y_{ij}^-$ are sub-Gaussian with parameter at most $c_2\sigma/\sqrt{n}$ for some absolute constant $c_2 > 0$. By Hoeffding's inequality for sums of sub-Gaussian random variables (see e.g. Vershynin (2012, Proposition 5.10)),

$$\mathbb{P}^{U_0,W_0}\biggl(\sum_{i\in S,j\in T} Z_{ij} > \frac{c_1\epsilon}{12\sqrt{n}}kk_1\biggr) \geq 1 - 2\exp\biggl\{-\frac{(\frac{c_1\epsilon}{12\sqrt{n}}kk_1)^2}{2c_2^2\sigma^2 k_1k_2/n}\biggr\} \to 1. \tag{3}$$

By (3) and the fact that $\mathbb{P}(\Omega_1) \to 1$, the event

$$\Omega_2 := \biggl\{\sum_{i\in S,j\in T} Z_{ij} \geq \frac{c_1\epsilon kk_1}{12\sqrt{n}}\biggr\}$$

has asymptotic probability 1.

Now define

$$\begin{aligned}\tilde{S} &= \{i \in \{1,\ldots,n\} : u_{an+i} \in U\cap K \text{ for some } 0 \leq a \leq \ell-1\}\\ \tilde{T} &= \{j \in \{1,\ldots,n\} : w_{bn+j} \in W\cap K \text{ for some } 0 \leq b \leq \ell-1\}\end{aligned}$$

Also, define $v^{(b)} = (v_{bn+1},\ldots,v_{bn+n})^\top$ for $0 \leq b \leq \ell-1$, $\tilde{v}_{\mathrm{sum}} = \sum_{0\leq b\leq \ell-1} v^{(b)}$ and $\tilde{v} = \tilde{v}_{\mathrm{sum}}/\|\tilde{v}_{\mathrm{sum}}\|_2$. By Lemma 6, we have $\|\tilde{v}_{\mathrm{sum}}\|_\infty \leq c_2k_1^{-1/2}$ with asymptotic probability 1 for some c_2 depending on β only. Hence $\|\tilde{v}_{\mathrm{sum}}\|_2 \leq c_2$. Thus, by Cauchy–Schwarz inequality, we have with asymptotic probability 1,

$$\|\tilde{X}_{\tilde{S}*}\tilde{v}\|_2 \geq \|\tilde{v}_{\mathrm{sum}}\|_2^{-1}(\#\tilde{S})^{-1/2}\|\tilde{X}_{\tilde{S}*}\tilde{v}_{\mathrm{sum}}\|_1$$

Since

$$\tilde{X}_{\tilde{S}*}\tilde{v}_{\mathrm{sum}} = \ell^{-1}\biggl(\sum_{0\leq a,b<\ell} Z_{S*}^{(a,b)}\biggr)\biggl(\sum_{0\leq b'<\ell} v^{(b)}\biggr) = \ell^{-1}\sum_{0\leq a,b<\ell} Z_{S*}^{(a,b)}v^{(b)} + \ell^{-1}\sum_{\substack{0\leq a,b,b'<\ell\\ b\neq b'}} Z_{S*}^{(a,b)}v^{(b')}$$

We can bound $\|\tilde{X}_{\tilde{S}*}\tilde{v}_{\mathrm{sum}}\|_1$ from below by the entrywise sums of the two terms above. The entrywise sum of the first term can be rewritten as $\ell^{-1}\sum_{i\in S,j\in T} Z_{ij}$, which by (3) is bounded from below by $\frac{c_3\epsilon k}{\ell\sqrt{n}}$ with asymptotic probability 1. The second term has entries with nonnegative means, hence another application of the Hoeffding's inequality shows that its contribution will be of smaller order than the first term with high probability. To summarise, we have that

$$\|\tilde{X}_{\tilde{S}*}\tilde{v}\|_2 \geq \frac{c_3\epsilon k}{\ell\sqrt{n}}.$$

with asymptotic probability 1. On the other hand, the submatrix $\tilde{X}_{\tilde{S}^c*}$ has independent and identically distributed entries. By Vershynin (2012, Lemma 5.9), for $i \in \tilde{S}^c$ and $1 \leq j \leq n$, $\tilde{X}_{ij} = \ell^{-1}\sum_{a,b=0}^{\ell-1} Z_{an+i,bn+j}^{(a,b)}$ is a centred sub-Gaussian random variable with sub-Gaussian parameter $\sigma/\sqrt{n}$ and variance $1/n$. Let $\tilde{X}_i$ denote the ith row vector of the matrix $\tilde{X}$, then conditional on $\tilde{T}$, we have that $\tilde{X}_i^\top\tilde{v}$ is also a centred sub-Gaussian random variable with parameter $\sigma/\sqrt{n}$ and variance $1/n$. Using Lemma 5, we have

$$\mathbb{P}\biggl(\|\tilde{X}_{S^c*}\tilde{v}\|_2^2 - \frac{n-\#\tilde{S}}{n} \leq -\sqrt{\frac{\log n}{n-\#\tilde{S}}}\biggr) \leq \exp\biggl\{-\frac{\log n}{64\sigma^4}\biggr\} \to 0.$$

Since $\#\tilde{S} \leq k_2$ with asymptotic probability 1, the event

$$\Omega_3 := \left\{ \|\tilde{X}_{\tilde{S}^{c}*}\tilde{v}\|_2^2 \geq 1 - \frac{k_2}{n} - \sqrt{\frac{2\log n}{n}} \right\}$$

has asymptotic probability 1. Finally, since $\tilde{X}\tilde{v} = (\tilde{X}_{\tilde{S}*}\tilde{v}, \tilde{X}_{\tilde{S}^{c}*}v)^\top$, on $\Omega_2 \cap \Omega_3$,

$$\|\tilde{X}\tilde{v}\|_2^2 = \|\tilde{X}_{\tilde{S}*}\tilde{v}\|_2^2 + \|\tilde{X}_{\tilde{S}^{c}*}v\|_2^2 \geq 1 + \frac{c_3^2\epsilon^2 k^2}{\ell^2 n} - \frac{k_2}{n} - \sqrt{\frac{2\log n}{n}}.$$

The right hand side is at least $1 + ck^2/(n\ell^2)$ for some absolute positive constant c for all large values of n. This verifies (1) and concludes the proof. □

Lemma 1. *Let G be a graph on m vertices and K a κ-subset of $V(G)$, such that the edge density of G restricted to K is at least $1/2 + \epsilon$. Let n, p be integers less than $m/2$. Choose $u_1, \ldots, u_n$ and $w_1, \ldots, w_p$ independently at random without replacement from $V(G)$. Denote $U = \{u_1, \ldots, u_n\}$ and $W = \{w_1, \ldots, w_p\}$. Define $N_{U,W;K}$ to be the number of edges with two endpoints in U and W respectively. Then for m, n, p, κ sufficiently large.*

$$\begin{aligned}
\mathbb{P}\left\{ \left| \#U \cap K - \frac{n\kappa}{m} \right| \geq \frac{\epsilon}{8}\frac{n\kappa}{m} \right\} &\leq \frac{64m}{\epsilon^2 n\kappa}, \\
\mathbb{P}\left\{ \left| \#W \cap K - \frac{p\kappa}{m} \right| \geq \frac{\epsilon}{8}\frac{p\kappa}{m} \right\} &\leq \frac{64m}{\epsilon^2 p\kappa}, \\
\mathbb{P}\left\{ N_{U,W;K} \leq \left(\frac{1}{2} + \frac{\epsilon}{4}\right)\frac{np\kappa^2}{m^2} \right\} &\leq \frac{16m(p\kappa + n\kappa + m)}{\epsilon^2 np\kappa^2}.
\end{aligned}$$

Proof. The cardinality of $U \cap K$ has $\mathrm{HyperGeom}(m, \kappa, n)$ distribution. Hence

$$\mathbb{E}(\#U \cap K) = \frac{n\kappa}{m} \quad \text{and} \quad \mathrm{var}(\#U \cap K) = n\frac{\kappa}{m}\frac{m-\kappa}{m}\frac{m-n}{m-1} \leq \frac{n\kappa}{m}.$$

The first inequality in the lemma now follows from an application of Chebyshev's inequality. A similar argument establishes the second inequality. For the final inequality in the lemma, we have that for κ sufficiently large,

$$\begin{aligned}
\mathbb{E}(N_{U,W;K}) &= \sum_{u\in K}\sum_{w\in K} \mathbb{P}(u \in U, w \in W)\mathbb{1}\{u \sim w\} \\
&= \frac{np}{m(m-1)}\sum_{u\in K}\sum_{w\in K}\mathbb{1}\{u \sim w\} \geq \left(\frac{1}{2} + \epsilon\right)\frac{np\kappa(\kappa-1)}{m(m-1)} \geq \left(\frac{1}{2} + \frac{\epsilon}{2}\right)\frac{np\kappa^2}{m^2}.
\end{aligned}$$

We then compute the variance of $N_{U,W;K}$ by

$$\begin{aligned}
\mathrm{var}(N_{U,W;K}) &= \mathrm{cov}\Big(\sum_{u\in K}\sum_{w\in K}\mathbb{1}\{u \in U, w \in W, u \sim w\}, \sum_{u'\in K}\sum_{w'\in K}\mathbb{1}\{u' \in U, w' \in W, u' \sim w'\}\Big) \\
&= \sum_{u,w,u',w'\in K} \mathrm{cov}\big(\mathbb{1}\{u \in U, w \in W, u \sim w\}, \mathbb{1}\{u' \in U, w' \in W, u' \sim w'\}\big) \\
&=: \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV},
\end{aligned}$$

where the four terms I, II, III and IV handle sums over subsets of indices $\{(u, w, u', w') \in K^4 : u \neq u', w \neq w'\}$, $\{(u, w, u', w') \in K^4 : u = u', w \neq w'\}$, $\{(u, w, u', w') \in K^4 : u \neq u', w = w'\}$ and $\{(u, w, u', w') \in K^4 : u = u', w = w'\}$ respectively.

We bound the four terms separately. For the first term, we have

$$\begin{aligned}
\mathrm{I} &= \sum_{u,u',w,w' \text{ distinct}} \Big\{\mathbb{P}(u, u' \in U, w, w' \in W) - \mathbb{P}(u \in U, w \in W)\mathbb{P}(u' \in U, w' \in W)\Big\}\mathbb{1}\{v \sim w\}\mathbb{1}\{u' \sim w'\} \\
&= \sum_{u,u',w,w' \text{ distinct}} \left\{\frac{n(n-1)p(p-1)}{m(m-1)(m-2)(m-3)} - \left(\frac{np}{m(m-1)}\right)^2\right\}\mathbb{1}\{u \sim w\}\mathbb{1}\{u' \sim w'\}.
\end{aligned}$$

When $m > \max(2n, 2p)$, the term in bracket above is non-positive, hence $\mathrm{I} \le 0$. For the second term, we get that

$$\begin{aligned}
\mathrm{II} &= \sum_{u,w,w' \text{ distinct}} \Big\{\mathbb{P}(u \in U, w, w' \in W) - \mathbb{P}(u \in U, w \in W)\mathbb{P}(u \in U, w' \in W)\Big\} \mathbb{1}\{u \sim w\}\mathbb{1}\{u' \sim w'\} \\
&= \sum_{u,w,w' \text{ distinct}} \bigg\{\frac{np(p-1)}{m(m-1)(m-2)} - \Big(\frac{np}{m(m-1)}\Big)^2\bigg\} \mathbb{1}\{u \sim w\}\mathbb{1}\{u \sim w'\} \\
&\le \frac{np(p-1)}{m(m-1)(m-2)} \sum_{u,w,w' \text{ distinct}} \mathbb{1}\{u \sim w\}\mathbb{1}\{u \sim w'\} \le \frac{np^2\kappa^3}{m^3}.
\end{aligned}$$

Similarly, we have

$$\mathrm{III} \le \frac{n(n-1)p\kappa(\kappa-1)(\kappa-2)}{m(m-1)(m-2)} \le \frac{n^2p\kappa^3}{m^3}.$$

And finally,

$$\mathrm{IV} = \sum_{u,w \text{ distinct}} \Big\{\mathbb{P}(u \in U, w \in W) - \mathbb{P}(u \in U, w \in W)^2\Big\} \mathbb{1}\{u \sim w\} \le \frac{np\kappa(\kappa-1)}{m(m-1)} \le \frac{np\kappa^2}{m^2}.$$

Sum up the four terms, we get that

$$\operatorname{var}(N_{U,W;K}) \le \frac{np\kappa^2}{m^2}\bigg(\frac{p\kappa}{m} + \frac{n\kappa}{m} + 1\bigg).$$

By Chebyshev's inequality, we get that

$$\mathbb{P}\bigg\{N_{U,W;K} \le \Big(\frac{1}{2} + \frac{\epsilon}{4}\Big)\frac{np\kappa^2}{m^2}\bigg\} \le \frac{16m(p\kappa + n\kappa + m)}{\epsilon^2 np\kappa^2},$$

as desired. □

B Auxiliary Results

Proof of Proposition 1. Let X_i denote the ith row vector of X. Then for any fixed $u \in \mathbb{S}^p(k)$,

$$\mathbb{E}e^{\lambda(X_i^\top u)} = \prod_{1\le j\le p} \mathbb{E}e^{\lambda X_{ij}u_j} \le \prod_j e^{\lambda^2 u_j^2/(2\sigma^2 n)} = e^{\lambda^2/(2\sigma^2 n)}.$$

Apply Lemma 5 to $\|Xu\|_2^2 - 1 = n^{-1}\sum_{i=1}^n \big\{(\sqrt{n}X_i^\top u)^2 - \mathbb{E}(\sqrt{n}X_i^\top u)^2\big\}$, and use the fact that $\theta/(8\sigma^2) \le 1$, we have

$$\mathbb{P}\big(1 - \theta \le \|Xu\|_2^2 \le 1 + \theta\big) \ge 1 - 2e^{-n\theta^2/(64\sigma^4)}.$$

We claim that there is a set $\mathcal{N}$ of cardinality at most $\binom{p}{k}9^k$ such that

$$\sup_{u\in\mathbb{S}^p(k)} \big|\|Xu\|_2^2 - 1\big| \le 2\sup_{u\in\mathcal{N}}\big|\|Xu\|_2^2 - 1\big| \tag{4}$$

Given (4), by union bound, we have

$$\begin{aligned}
\mathbb{P}(X \in \mathrm{RIP}(k,\theta)) &= \mathbb{P}\Big(\sup_{u\in\mathbb{S}^p(k)} \big|\|Xu\|_2^2 - 1\big| \le \theta\Big) \ge \mathbb{P}\Big(\sup_{u\in\mathcal{N}}\big|\|Xu\|_2^2 - 1\big| \le \theta/2\Big) \\
&\ge 1 - 2\binom{p}{k}9^k e^{-n\theta^2/(256\sigma^4)} \ge 1 - 2\exp\bigg\{k\log\Big(\frac{9ep}{k}\Big) - \frac{n\theta^2}{256\sigma^4}\bigg\},
\end{aligned}$$

as desired. It remains to verify Claim (4). For any cardinality k subset $J \subseteq \{1, \ldots, p\}$, let $B_J = \{u \in \mathbb{S}^p(k) : u_{J^c} = 0\}$. Each B_J contains a $1/4$-net, $\mathcal{N}_J$, of cardinality at most 9^k (Vershynin, 2012, Lemma 5.2). Then $\mathcal{N} := \cup_J \mathcal{N}_J$ form a $1/4$-net for $\mathbb{S}^p(k)$. Define $u_J \in \operatorname{argmax}_{u\in B_J}\|Xu\|^2$

and let v_J be an element in $\mathcal{N}_J$ closest in Euclidean distance to u_J. Define $A := X^\top X - I_p$. We have

$$|u_J^\top A u_J| \le |v_J^\top A v_J| + |(u_J - v_J)^\top A v_J| + |u_J^\top A(u_J - v_J)| \le \max_{u\in\mathcal{N}_I} |u^\top A u| + \frac{1}{2}\sup_{u\in\mathbb{S}^p(k)} |u^\top A u|.$$

Hence

$$\sup_{u\in\mathbb{S}^p(k)} |u^\top A u| \le 2\max_{u\in\mathcal{N}} |u^\top A u|,$$

which verifies the claim. □

Proof of Proposition 2. By definition, $\|X^\top X - I_p\|_{\mathrm{op},k} \le \theta$ is equivalent to $X \in \mathrm{RIP}_{n,p}(k,\theta)$. Moreover, by Proposition 1, $X \in \mathrm{RIP}_{n,p}(k,\theta)$ with probability converging to 1, under $\tilde{Q}^{\otimes(n\times p)}$. The certifier hence satisfies the two desired properties. □

Proof of Proposition 3. The proposed certifier is clearly polynomial time computable (it has time complexity $O(np^2)$). To verify that it is a certifier, we check that (i) $\psi_n^{-1}(1) \subseteq \mathrm{RIP}_{n,p}(k,\theta)$ and (ii) $\liminf_{n\to\infty} \tilde{Q}^{\otimes(n\times p)}(\psi_n^{-1}(1)) > 2/3$.

For (i), on the event $\|X^\top X - I_p\|_\infty \le 14\sigma^2\sqrt{\frac{\log p}{n}}$, for any index set $T \in \{1,\ldots,p\}$ of cardinality k, we have $\|X_{*T}^\top X_{*T} - I_k\|_\infty \le 14\sigma^2\sqrt{\frac{\log p}{n}}$, which implies that

$$\|X_{*T}^\top X_{*T} - I_k\|_{\mathrm{op}} \le 14\sigma^2 k\sqrt{\frac{\log p}{n}} \le \theta$$

For (ii), let $Y_n \sim \chi_n^2$. Using Lemma 5 and the fact that for any $A \in \mathbb{R}^{p\times p}$

$$\|A\|_\infty = \sup_{S\subseteq\{1,\ldots,p\},\#S=2} \|A_{SS}\|_\infty \le \sup_{S\subseteq\{1,\ldots,p\},\#S=2} \|A_{SS}\|_{\mathrm{op}} = \|A\|_{\mathrm{op},2}$$

we get

$$\begin{aligned}
&\mathbb{P}\biggl\{\|X^\top X - I_p\|_\infty \le 14\sigma^2\sqrt{\frac{\log p}{n}}\biggr\} \ge \mathbb{P}\biggl\{\sup_{u\in\mathbb{S}^p(2)} \bigl|\|Xu\|_2^2 - 1\bigr| \le 14\sigma^2\sqrt{\frac{\log p}{n}}\biggr\}\\
&\ge 1 - 2\binom{p}{2}9^2\exp\biggl\{-\frac{n}{256\sigma^4}\frac{196\sigma^4\log p}{n}\biggr\}\\
&\ge 1 - 81p^2\exp\{-3\log p/4\} \to 1.
\end{aligned}$$

as desired. □

Lemma 2. *Let Z be a non-negative random variable and $r \ge 2$, then*

$$\mathbb{E}(Z^r) \ge \mathbb{E}(|Z - \mathbb{E}Z|^r).$$

In other words, centring a nonnegative random variable shrinks its second or higher absolute moments.

Proof. Let $\mu := \mathbb{E}(Z)$ and define $Y = Z - \mu$. Let P denote the probability measure on $\mathbb{R}$ associated with random variable Y. Hence $\int_{[-\mu,\infty)} y\,dP(y) = 0$. Without loss of generality, we may assume that Z is not a point mass. Then $\int_{[-\mu,0]}(-y)\,dP(y) = \int_{(0,\infty)} y\,dP(y) = A$ for some $A > 0$. For any measureable function $f : \mathbb{R} \to [0,\infty)$, we may write

$$\begin{aligned}
A\int_{[-\mu,\infty)} f(y)\,dP(y) &= \int_{[-\mu,0]}(-v)\,dP(v)\int_{(0,\infty)} f(u)\,dP(u) + \int_{(0,\infty)} u\,dP(u)\int_{[-\mu,0]} f(v)dP(v)\\
&= \int_{u\in(0,\infty)}\int_{v\in[-\mu,0]}\biggl(\frac{u}{u-v}f(v) - \frac{v}{u-v}f(u)\biggr)(u-v)\,dP(v)\,dP(u).
\end{aligned} \tag{5}$$

Let (U,V) be a bivariate random vector having probability measure

$$\frac{1}{A}(u-v)\mathbb{1}_{(0,\infty)}(u)\mathbb{1}_{[-\mu,0]}(v)\,dP(u)\,dP(v)$$

on $\mathbb{R}^2$ (that this is a probability measure follows from substituting $f(y)\equiv 1$ in (5)). Then (5) can be rewritten as

$$\mathbb{E}\{f(Y)\} = \mathbb{E}\biggl\{\frac{U}{U-V}f(V) - \frac{V}{U-V}f(U)\biggr\}.$$

Now consider choosing f to be $f_1(y) = |y|^r$ and $f_2(y) = (y+\mu)^r$ respectively in the above equation. Note that for $u\in(0,\infty)$ and $v\in[-\mu,0]$ and $r\geq 2$, we always have

$$uf_2(v) - vf_2(u) \geq -vf_2(u) \geq -v(u-v)^r \geq (-v)^r u + (-v)u^r \geq uf_1(v) - vf_1(u).$$

Therefore,

$$\begin{aligned}\mathbb{E}(|Y|^m) &= \mathbb{E}\biggl\{\frac{U}{U-V}f_1(V) - \frac{V}{U-V}f_1(U)\biggr\}\\ &\leq \mathbb{E}\biggl\{\frac{U}{U-V}f_2(V) - \frac{V}{U-V}f_2(U)\biggr\} = \mathbb{E}(|Y+\mu|^m),\end{aligned}$$

as desired. □

Lemma 3. *Suppose X is a sub-Gaussian random variable with parameter σ and median ξ. Let $X^+ = X \mid X\geq\xi$ and $X^- = X \mid X<\xi$. Then $X^+ - \mathbb{E}X^+$ and $X^- - \mathbb{E}X^-$ are both sub-Gaussian with parameters are most $c\sigma$ for some absolute constant c.*

Proof. By Vershynin (2012, Lemma 5.5), X is sub-Gaussian with parameter σ implies that $(\mathbb{E}|X|^p)^{1/p}\leq c_1\sigma\sqrt{p}$ for some absolute constant c_1. Hence by Lemma 2, we have

$$\mathbb{E}\bigl(\bigl|X^+ - \mathbb{E}X^+\bigr|^p\bigr)^{1/p} \leq \bigl(\mathbb{E}\bigl|X^+\bigr|^p\bigr)^{1/p} = 2\bigl(\mathbb{E}\bigl|X\mathbb{1}\{X\geq\xi\}\bigr|^p\bigr)^{1/p} \leq 2c_1\sigma\sqrt{p}.$$

Using Vershynin (2012, Lemma 5.5) again, we have that $X^+ - \mathbb{E}X^+$ is sub-Gaussian with parameter at most $c\sigma$ for some absolute constant c. A similar argument holds for $X^- - \mathbb{E}X^-$. □

Lemma 4. *Suppose X is a random variable satisfying $\mathbb{E}e^{\lambda X}\leq e^{\sigma^2\lambda^2/2}$ for all $\lambda\in\mathbb{R}$. Define $Y = X^2 - \mathbb{E}X^2$. Then $\mathbb{E}e^{\lambda Y}\leq e^{16\sigma^4\lambda^2}$ for all $|\lambda|\leq\frac{1}{4\sigma^2}$.*

Proof. By Markov's inequality,

$$\mathbb{P}(|X|\geq t) = \mathbb{P}(X\geq t) + \mathbb{P}(-X\geq t) \leq e^{-t^2/\sigma^2}\mathbb{E}\bigl(e^{tX/\sigma^2}\bigr) + e^{-t^2/\sigma^2}\mathbb{E}\bigl(e^{-tX/\sigma^2}\bigr) \leq 2e^{-t^2/(2\sigma^2)}.$$

From Lemma 2, for $r\geq 2$

$$\mathbb{E}(|Y|^r) \leq \mathbb{E}(|X|^{2r}) = \int_0^\infty \mathbb{P}(|X|\geq t)(2r)t^{2r-1}\,dt \leq \int_0^\infty 4rt^{2r-1}e^{-t^2/(2\sigma^2)}\,dt = 2(2\sigma^2)^r\Gamma(r+1).$$

Consequently, if $|2\sigma^2\lambda|\leq 1/2$, then

$$\mathbb{E}e^{\lambda Y} = \sum_{r=0}^\infty \frac{\lambda^r\mathbb{E}Y^r}{r!} \leq 1 + 2\sum_{r=2}^\infty (2\sigma^2\lambda)^r \leq 1 + 16\sigma^4\lambda^2 \leq e^{16\sigma^4\lambda^2},$$

as desired. □

Lemma 5. *Let $X_1, X_2, \ldots, X_n$ be independent sub-Gaussian random variables with sub-Gaussian parameters at most σ. Let $Y_i := X_i^2 - \mathbb{E}X_i^2$. Then*

$$\mathbb{P}\biggl(\sum_{i=1}^n Y_i \geq \theta\biggr) \leq \exp\biggl\{-\biggl(\frac{\theta^2}{64n\sigma^4}\wedge\frac{\theta}{8\sigma^2}\biggr)\biggr\}$$

$$\mathbb{P}\biggl(\sum_{i=1}^n Y_i \leq -\theta\biggr) \leq \exp\biggl\{-\frac{\theta^2}{64n\sigma^4}\biggr\}$$

Proof. Using Markov's inequality, we have

$$\mathbb{P}\Big(\sum_{i=1}^{n} Y_i \geq \theta\Big) = \mathbb{P}\Big(e^{\lambda \sum_i Y_i} \geq e^{\lambda\theta}\Big) \leq e^{-\lambda\theta} \prod_i \mathbb{E}e^{\lambda Y_i}.$$

Set $\lambda = \frac{\theta}{32n\sigma^4} \wedge \frac{1}{4\sigma^2}$. By Lemma 4, we have

$$\mathbb{P}\Big(\sum_{i=1}^{n} Y_i \geq \theta\Big) \leq e^{-\lambda\theta + 16\lambda^2 n\sigma^4} \leq e^{-\lambda\theta/2},$$

which establishes the first desired inequality. Applying the same argument with $-Y_i$ in place of Y_i we get

$$\mathbb{P}\Big(\sum_{i=1}^{n} Y_i \leq -\theta\Big) \leq \exp\Big\{-\Big(\frac{\theta^2}{64n\sigma^4} \wedge \frac{\theta}{8\sigma^2}\Big)\Big\}. \tag{6}$$

Taylor expand the moment generating function of X_i around 0, we have $\mathbb{E}X_i^2 \leq \sigma^2$. Hence we may assume $\theta \leq n\sigma^2$. Then we have

$$\frac{\theta^2}{64n\sigma^4} < \frac{\theta}{8\sigma^2},$$

which together with (6) implies the desired result. □

Lemma 6. *Suppose $n\ell$ balls are arranged in an array of n rows and ℓ columns and k balls ($k < n$) are chosen uniformly at random. Let V_i be the number of chosen balls in row i and $V = (V_1, \ldots, V_n)^\top$. Then*

$$\mathbb{P}\Big(\|V\|_0 \leq k - \frac{k^2}{2n} - \sqrt{k\log k}\Big) \leq \frac{1}{k^2}.$$

Moreover, if $k \leq n^\gamma$ for some $\gamma < 1$, then

$$\mathbb{P}\big(\|V\|_\infty \geq a\big) \leq n^{1-a(1-\gamma)}\big(1 - n^{-(1-\gamma)}\big).$$

Proof. Let U_i be the number of balls chosen in row i when balls are drawn with replacement from the array and $U = (U_1, \ldots, U_n)^\top$. Then $\|V\|_0$ is stochastically larger than $\|U\|_0$ and $\|V\|_\infty$ is stochastically smaller than $\|U\|_\infty$. So it suffices to show the desired inequalities with U replacing V. In the following argument, we consider only drawing with replacement.

Let $\mathcal{X} = \{e_1, \ldots, e_n\}$ where e_i denotes the ith standard basis vector in $\mathbb{R}^n$. For $1 \leq r \leq k$, let X_r be uniformly distributed in $\mathcal{X}$. Then $U \overset{d}{=} \sum_{r=1}^{k} X_r$. We note that changing the value of any of the X_r affects the value of $\|U\|_0$ by at most 1. By McDiarmid's inequality (McDiarmid, 1989), we have that for any $t > 0$,

$$\mathbb{P}\big(\|U\|_0 - \mathbb{E}\|U\|_0 \leq -t\big) \leq e^{-\frac{2t^2}{k}}. \tag{7}$$

For $1 \leq i \leq n$. Define $J_i = \mathbb{1}\{\text{no ball is chosen in row } i\}$, then

$$\mathbb{E}\|U\|_0 = n - \sum_{i=1}^{n} \mathbb{E}J_i = n - n(1 - 1/n)^k \geq k\Big(1 - \frac{k}{2n}\Big).$$

Thus, together with (7), we have

$$\mathbb{P}\Big(\|U\|_0 \leq k - \frac{k^2}{2n} - \sqrt{k\log k}\Big) \leq \mathbb{P}\Big(\|U\|_0 - \mathbb{E}\|U\|_0 \leq -\sqrt{k\log k}\Big) \leq e^{-2\log k} = k^{-2},$$

as desired. For the second inequality,

we have by union bound that

$$\begin{aligned}\mathbb{P}(\|U\|_\infty \geq a) \leq n\mathbb{P}(U_1 \geq a) &= n\sum_{s=a}^{k}\binom{k}{s}n^{-s} \\ &\leq n\sum_{s=a}^{\infty}(k/n)^s = n\frac{(k/n)^a}{1-k/n} \leq n^{1-a(1-\gamma)}(1 - n^{-(1-\gamma)})^{-1},\end{aligned}$$

as desired. □

References

McDiarmid, C. (1989) On the method of bounded differences. *Surveys in Combinatorics*, **141**, 148–188.

Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.) *Compressed Sensing, Theory and Applications*. Cambridge University Press, Cambridge. 210–268.