The Annals of Statistics
2015, Vol. 43, No. 3, 1089–1116
DOI: 10.1214/14-AOS1300
© Institute of Mathematical Statistics, 2015

COMPUTATIONAL BARRIERS IN MINIMAX SUBMATRIX DETECTION

BY ZONGMING MA[1] AND YIHONG WU

University of Pennsylvania and University of Illinois at Urbana-Champaign

This paper studies the minimax detection of a small submatrix of elevated mean in a large matrix contaminated by additive Gaussian noise. To investigate the tradeoff between statistical performance and computational cost from a complexity-theoretic perspective, we consider a sequence of discretized models which are asymptotically equivalent to the Gaussian model. Under the hypothesis that the planted clique detection problem cannot be solved in randomized polynomial time when the clique size is of smaller order than the square root of the graph size, the following phase transition phenomenon is established: when the size of the large matrix $p \to \infty$, if the submatrix size $k = \Theta(p^{\alpha})$ for any $\alpha \in (0, 2/3)$, computational complexity constraints can incur a severe penalty on the statistical performance in the sense that any randomized polynomial-time test is minimax suboptimal by a polynomial factor in p; if $k = \Theta(p^{\alpha})$ for any $\alpha \in (2/3, 1)$, minimax optimal detection can be attained within constant factors in linear time. Using Schatten norm loss as a representative example, we show that the hardness of attaining the minimax estimation rate can crucially depend on the loss function. Implications on the hardness of support recovery are also obtained.

1. Introduction.

Statistical inference of structured large matrices lies at the heart of many applications involving massive datasets, such as matrix completion, functional genomics, community detection and clustering; see, for instance, [6, 13, 14, 38, 42] and the references therein. Many of these detection and estimation problems have been investigated from a decision theoretic viewpoint, where one first establishes a minimax lower bound for any test or estimator and then constructs a specific procedure which attains the lower bound within a constant or logarithmic factor.

An important element absent from the foregoing decision theoretic paradigm is computational complexity. This aspect is especially relevant in the context of high-dimensional statistical inference, where computationally efficient procedures (e.g., convex programming, iterative algorithms, etc.) are highly desirable. However, it has been empirically observed in several basic detection and estimation problems

Received August 2014; revised December 2014.

[1] Supported in part by NSF Career Award DMS-13-52060 and the Claude Marion Endowed Faculty Scholar Award of the Wharton School.

MSC2010 subject classifications. Primary 62H15; secondary 62C20.

Key words and phrases. Asymptotic equivalence, high-dimensional statistics, computational complexity, minimax rate, planted clique, submatrix detection.

that popular low-complexity algorithms fail to attain the minimax rates; see, for example, [6, 8–10, 13, 30]. This invites the following question: how much do we need to back off from the statistical optimality due to computational complexity constraints? In this paper, we revisit the sparse submatrix detection problem that has been studied in [8, 11, 13, 28, 38, 41], where the goal is to detect a small submatrix with elevated mean in a large noisy matrix. Motivations for this detection problem include biclustering for analyzing microarray data [38] and community detection in social networks [6], etc.

1.1. *Problem formulation.* Let $X = (X_{ij})$ be a $p \times p$ matrix with independent Gaussian entries $X_{ij} \overset{\text{ind.}}{\sim} N(\theta_{ij}, 1)$. Denote the mean matrix by $\theta = (\theta_{ij}) \in \mathbb{R}^{p\times p}$ and the distribution of X by $\mathbb{P}_\theta$. The submatrix detection problem deals with the following setup [13]: under the null hypothesis, the signal is absent, and θ is a zero matrix. Under the alternative hypothesis, θ is zero except for a submatrix of size at least $k \times k$ where all the entries exceed some positive value λ. In other words, detecting the submatrix boils down to testing the following hypotheses on the mean matrix:

$$H_0 : X \sim \mathbb{P}_0 \quad \text{versus} \quad H_1 : X \sim \mathbb{P}_\theta, \qquad \theta \in \mathcal{M}(p,k,\lambda), \tag{1}$$

where $\mathbb{P}_0$ is standard Gaussian, and the parameter space for the alternative hypothesis is

$$\begin{aligned} \mathcal{M}(p,k,\lambda) = \bigl\{\theta \in \mathbb{R}^{p\times p} : \exists U, V \subset [p],\ \text{s.t. } |U|, |V| \geq k, \\ \theta_{ij} \geq \lambda,\ \text{if } (i,j) \in U \times V, \theta_{ij} = 0 \text{ if } (i,j) \notin U \times V\bigr\}. \end{aligned} \tag{2}$$

In this problem, the key parameters are the matrix dimension p, the block size k and the signal magnitude λ. Clearly, it is easier to detect the submatrix if either k or λ increases. Throughout the paper, we focus on the asymptotic setting where p tends to infinity and both $k = k(p)$ and $\lambda = \lambda(p)$ are functions of p, though we typically drop the explicit dependence on p for conciseness.

For any test $\phi : \mathbb{R}^{p\times p} \to \{0, 1\}$, we denote its worst-case Type-I + II error probability of testing (1) by

$$\mathcal{E}(\phi) = \mathbb{P}_0\bigl\{\phi(X) = 1\bigr\} + \sup_{\theta\in\mathcal{M}(p,k,\lambda)} \mathbb{P}_\theta\bigl\{\phi(X) = 0\bigr\}. \tag{3}$$

The optimal total probability of error is denoted by

$$\mathcal{E}^* = \inf_{\phi : \mathbb{R}^{p\times p}\to\{0,1\}} \mathcal{E}(\phi). \tag{4}$$

In the asymptotic regime of

$$p \to \infty, \qquad k \to \infty \quad \text{and} \quad k/p \to 0, \tag{5}$$

the necessary and sufficient condition for reliably detecting the submatrix has been characterized by Butucea and Ingster ([13], Theorems 2.1 and 2.2): $\mathcal{E}^* \to 0$ if

$$\frac{\lambda}{p/k^2} \to \infty \quad \text{or} \quad \liminf_{p\to\infty} \frac{\lambda}{2\sqrt{(1/k)\log(p/k)}} > 1, \tag{6}$$

and, conversely, $\mathcal{E}^* \to 1$ if

$$\frac{\lambda}{p/k^2} \to 0 \quad \text{and} \quad \limsup_{p\to\infty} \frac{\lambda}{2\sqrt{(1/k)\log(p/k)}} < 1. \tag{7}$$

From this point forward, we say reliable detection is statistically possible if $\mathcal{E}^* \to 0$ and a sequence of tests $\{\phi_p\}$ reliably detects the submatrix if $\mathcal{E}(\phi_p) \to 0$.

To reliably detect the submatrix under condition (6), Butucea and Ingster [13] proposed a test involving enumerating all $k \times k$ submatrices of X, which is asymptotically optimal but computationally intensive. It is unclear from first principles whether statistically optimal detection can be achieved using computationally efficient procedures. Thus an intriguing question is in order: under the optimal condition (6) so that $\mathcal{E}^* \to 0$, is there a sequence of computationally efficient tests $\{\phi_p\}$ such that $\mathcal{E}(\phi_p) \to 0$?

1.2. *The penalty incurred by complexity constraints.* To approach the computational hardness of the submatrix detection problem rigorously, we need to investigate the computational cost of testing procedures in a complexity theoretic sense. However, an immediate hurdle for the Gaussian model (1) is that computational complexity is not well defined for all tests dealing with nondiscrete distributions since the observation cannot be represented by finitely many bits almost surely. To propose a paradigm for complexity-constrained hypotheses testing, we consider a sequence of *discretized* Gaussian models which is asymptotically equivalent to the original model in the sense of Le Cam [33] and hence preserves the statistical difficulty of the problem. More importantly, the computational complexity of tests on the discretized model can be appropriately defined. See Section 3 for details.

Next, we take the standard reduction approach in complexity theory: we show that if the signal magnitude is smaller than a certain threshold, detecting the submatrix is computationally no easier than certain well-known intractable problems. In other words, if an efficient method existed for submatrix detection, it would lead to an efficient solution to this problem. In this paper, we use the *planted clique* problem as the benchmark, which deals with detecting whether a given instance of an Erdős–Rényi random graph of size N contains a planted clique of size κ. It is widely believed that the detection problem cannot be solved in randomized polynomial time when $\kappa = o(\sqrt{N})$, which we shall refer to as the *planted clique hypothesis*. For the precise statement and further discussions, see Definition 1 and Hypothesis 1 in Section 4.

Assuming the planted clique hypothesis, our main finding (Theorem 2 in Section 4) characterizes when it is possible to achieve reliable detection using computationally efficient procedures and when it is impossible. The core of the arguments lies in a *randomized polynomial-time reduction* scheme which maps the $N \times N$ adjacency matrix of the random graph in the planted clique problem to a $p \times p$ random matrix in polynomial time. It is worth noting that when $k \geq p^{\alpha}$ for some $\alpha \geq \frac{1}{2}$, the cardinality of the graph N is not equal to the size of the matrix p but rather chosen to be $p^{1+\delta}$ (omitting log factor), where $\delta > 0$ depends on α. On the other hand, κ can always be chosen as a constant multiple of k.

Our main result can be illustrated by focusing on the following asymptotic regime, where the submatrix size grows according to $k = \Theta(p^{\alpha})$, and the signal magnitude decays as $\lambda = \Theta(p^{-\beta})$ for fixed constants $\alpha \in (0, 1)$ and $\beta \in [0, 1]$[2] as $p \to \infty$. For any two numbers a and b, let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Define

$$\beta^* \triangleq \frac{\alpha}{2} \vee (2\alpha - 1) \geq \beta^\sharp \triangleq 0 \vee (2\alpha - 1).$$

The statistical and computational feasibility of the submatrix detection problem is demonstrated in Figure 1, where the (α, β)-plane is divided into three regions:

(1) $\beta > \beta^*$ (top region): reliable detection of the submatrix is statistically impossible because the signal is too weak.

(2) $\beta < \beta^\sharp$ (right triangular region): reliable submatrix detection is achievable by computationally efficient tests.

(3) $\beta^\sharp < \beta < \beta^*$ (lower left triangular region): reliable detection is statistically possible but computationally intractable, in the sense that it is at least as hard

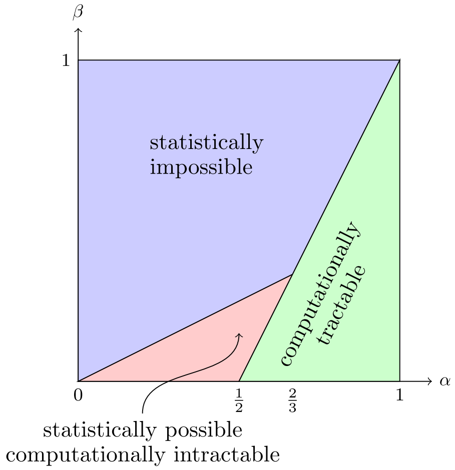

FIG. 1. *Detection boundary β^* versus efficiently computable detection boundary $\beta^\sharp$.*

[2] The regime of $\beta > 1$ is not interesting since the hypotheses are indistinguishable even if the submatrix becomes the whole matrix ($k = p$).

as solving the planted clique problem of a particular configuration, which is intractable under the planted clique hypothesis.

Therefore, the tractability of the submatrix detection problem undergoes a sharp transition: in the moderately sparse regime where $\alpha \in (2/3, 1)$, computational constraints incur no penalty on the statistical performance. In contrast, in the highly sparse regime where $\alpha \in (0, 2/3)$, achieving the statistical optimal boundary requires computational resources that are powerful enough to solve the planted clique problem, and, consequently, computationally efficient procedures require significantly higher signal-to-noise ratio to detect the submatrix.

The complexity-theoretic limits for the submatrix detection problem also lead to interesting findings for the related support recovery problem when the signal submatrix is present [8, 28, 38]. Moreover, it also sheds light on the statistical and computational tradeoff in the problem of estimating block sparse matrices [34]. In particular, we show the surprising result that the hardness of minimax estimation can crucially depend on the loss function, in the sense that attaining the minimax estimation rate can be computationally easy for one type of loss functions but hard for the other.

1.3. *Related works.* Despite the vast body of literature on developing computationally efficient procedures with optimal statistical performance for problems such as compressed sensing, rigorous results on inferential limits of statistical problems under computational complexity constraints are comparatively limited. A representative work is the investigation of the complexity of detecting sparse principal components by Berthet and Rigollet [9], which is one of the motivations of the present paper. Sparse principal component detection refers to the problem of testing $N(0, I_p)$ against $N(0, I_p + avv')$ for k-sparse unit vector v based on n i.i.d. observations [10]. Since the model is not discrete, as previously mentioned, the difficulty of ill-defined computational complexity is also present. In [9], the authors relaxed the Gaussian detection problem to a composite testing problem that includes discrete distributions, where the empirical projection variances of the null and alternative hypotheses satisfy respective uniform χ^2-tail type concentration inequalities. In the regime of $p^\delta \le k \le p^\alpha$ for some absolute constants $0 < \delta \le \alpha < \frac{1}{2}$ and $n \le p$, they showed that the computable detection rate for the deviation of the largest principal component from the rest of the spectrum is no smaller than $\sqrt{\frac{k^b}{n}}$ for any $b < 2$, which far exceeds the minimax detection rate of $\sqrt{\frac{k}{n}\log p}$.

Although both the authors of current paper and Berthet and Rigollet [9] use the planted clique hypothesis for studying complexity theoretic lower bounds, there are a few important differences. First, Berthet and Rigollet [9] extend the original "simple vs. composite" Gaussian sparse principal component detection problem into a "composite versus composite" testing problem, and the data no longer needs to be Gaussian. As a consequence, more distributions are included in both the null and alternative hypotheses, and thus constructing the reduction scheme becomes easier than for the original Gaussian hypotheses. In contrast, the current paper considers an asymptotically equivalent discretized model which is faithful to the original Gaussian submatrix detection problem in [13]. Second, the computational lower bounds in [9] are established only when the sparsity level satisfies $p^\delta \le k \le p^\alpha$ for $0 < \delta \le \alpha < \frac{1}{2}$. In comparison, due to a new reduction scheme, the current paper provides a more complete characterization of the computational limits for all $k \ge p^\delta$ and any $\delta > 0$. Last but not least, we propose an asymptotic equivalence framework in the sense of Le Cam, which preserves the statistical nature of the problem and, at the same time, allows rigorous statements of computational complexity of testing procedures. The approach via asymptotically equivalent discretized experiments is potentially useful in future works dealing with nondiscrete distributions.

In addition, some researchers have studied the minimax sub-optimality of certain computationally efficient methodologies, such as those based on convex relaxations, in an array of problems including estimating sparse eigenvectors [30], support recovery for sparse submatrices [8], combinatorial testing [1], community detection [6], etc. In some of the papers, the authors also conjecture that the minimax rate optimality cannot be achieved by any computationally efficient algorithms. From a different viewpoint, Chandrasekaran and Jordan [15] consider the tradeoff between computation and statistical performance within a specific family of algorithms parameterized by the level of convex relaxations in the classical normal mean estimation problem. In contrast, the goal of the present paper is to investigate the impact of complexity constraint on *any* statistical procedure for the submatrix detection problem.

1.4. *Organization of the paper.* The rest of the paper is organized as follows. In Section 2, we study test statistics for submatrix detection under Gaussian models. To incorporate computational complexity into the decision theoretic problem, we introduce in Section 3 a sequence of asymptotically equivalent discretized models and show that the minimax detection results (6)–(7) remain unchanged under these models. In Section 4, we state our main result in Theorem 2 under the planted clique hypothesis and present its proof with a concrete randomized polynomial-time procedure that reduces the planted clique problem to a Bayesian version of the submatrix detection problem. We discuss some related problems in Section 5. Section 6 presents additional proofs for results in earlier sections.

1.5. *Notation.* For any positive integer n, let $[n]$ denote the set $\{1, \dots, n\}$. For any $a \in \mathbb{R}$, let $a_+ = a \vee 0$. For any square matrix A, $\mathsf{Tr}(A) = \sum_i A_{ii}$ stands for its trace. For any two matrices A and B of the same size, $A \circ B$ denotes their component-wise product; that is, $(A \circ B)_{ij} = A_{ij}B_{ij}$, and $\langle A, B\rangle = \mathsf{Tr}(A'B)$. Let $\mathcal{L}(Y)$ denote the law, that is, the probability distribution, of a random variable Y.

Let $\mathcal{L}(Y|E)$ denote the distribution of Y conditioned on the event E. The total variation distance between distributions P and Q is $\mathsf{TV}(P, Q) \triangleq 1 - \int(\mathrm{d}P \wedge \mathrm{d}Q)$. For ease of notation, we also write $\mathsf{TV}(X, Y)$ in place of $\mathsf{TV}(\mathcal{L}(X), \mathcal{L}(Y))$ for random variables X and Y. We write $X \overset{(\mathrm{d})}{=} Y$ if $\mathcal{L}(X) = \mathcal{L}(Y)$. Let Φ, $\overline{\Phi} = 1 - \Phi$ and φ denote the distribution, survival and the probability density functions of the standard Gaussian distribution. For any set I, $|I|$ denotes its cardinality. For any sequences $\{a_p\}$ and $\{b_p\}$, we write $a_p \asymp b_p$ or $a_p = \Theta(b_p)$ if there is an absolute constant $C > 0$ such that $1/C \le a_p/b_p \le C$. We also write $a_p \ll b_p$ and $b_p \gg a_p$ if $a_p = o(b_p)$, and $a_p = \Omega(b_p)$ if $b_p = O(a_p)$.

2. Test statistics for submatrix detection. To prepare for later investigation, we first study three test statistics for the submatrix detection problem (1)–(2). The first two are the linear and the scan test statistics proposed in [13],

$$
\begin{aligned}
T_{\mathrm{lin}} &= T_{\mathrm{lin}}(X) \triangleq \frac{1}{p} \sum_{i,j=1}^{p} X_{ij}, \\
T_{\mathrm{scan}} &= T_{\mathrm{scan}}(X) \triangleq \frac{1}{k} \max_{|S|=|T|=k} \sum_{i\in S, j\in T} X_{ij}.
\end{aligned} \tag{8}
$$

In addition, we also consider the maximum test statistic

$$T_{\max} = T_{\max}(X) \triangleq \max_{i,j\in[p]} X_{ij}. \tag{9}$$

The following lemma gives nonasymptotic bounds on the Type-I + II error probabilities on tests based on these statistics. Recall the definition of $\mathcal{M}(p, k, \lambda)$ in (2).

LEMMA 1. *Let $\mathcal{M} = \mathcal{M}(p, k, \lambda)$ and $c > 0$ be any absolute constant. For T_{lin} in (8), set $\tau = \frac{\lambda k^2}{2p}$. Then*

$$\mathbb{P}_0\{T_{\mathrm{lin}} > \tau\} + \sup_{\theta\in\mathcal{M}} \mathbb{P}_\theta\{T_{\mathrm{lin}} \le \tau\} \le \mathrm{e}^{-\lambda^2 k^4/8p^2}. \tag{10}$$

For T_{scan} in (8), set $\tau' = \sqrt{(4 + c)\log\binom{p}{k}}$. Then

$$\mathbb{P}_0\{T_{\mathrm{scan}} > \tau'\} + \sup_{\theta\in\mathcal{M}} \mathbb{P}_\theta\{T_{\mathrm{scan}} \le \tau'\} \le \binom{p}{k}^{-c/2} + \mathrm{e}^{-(1/2)(\lambda k - \tau')_+^2}. \tag{11}$$

For $T_{\max}$ in (9), set $\tau'' = \sqrt{(4 + c)\log p}$. Then

$$\mathbb{P}_0\{T_{\max} > \tau''\} + \sup_{\theta\in\mathcal{M}} \mathbb{P}_\theta\{T_{\max} \le \tau''\} \le p^{-c/2} + \mathrm{e}^{-(1/2)(\lambda - \tau'')_+^2}. \tag{12}$$

For the proof of Lemma 1, see Section 6.1. By Lemma 1, we have $\mathcal{E}(\mathbf{1}_{\{T_{\mathrm{lin}}>\tau\}}) \to 0$ when the first condition in (6) holds, while $\mathcal{E}(\mathbf{1}_{\{T_{\mathrm{scan}}>\tau'\}}) \to 0$ when the second condition in (6) holds if we pick the constant c in τ' to be sufficiently small such that $\liminf_{p\to\infty} \lambda k/\tau' > 1$. The error bounds on $T_{\max}$ will be used later to establish the achievability part of the main result.

3. Asymptotically equivalent discretized model.

Gaussian distributions serve as good statistical models for many real-world datasets. However, as an idealized approximation, Gaussian experiment does not capture the finite-precision nature of statistical computing systems in reality. As mentioned in Section 1, it is an ill-defined problem to investigate the computational complexity of testing the Gaussian hypothesis (1) since the data do not admit any representation using finite bits. Therefore a new paradigm is needed in order to make sense of hypothesis testing with complexity constraints in general. There are two goals of the paradigm:

(a) to provide a rigorous framework for quantifying the complexity of statistical inference involving continuous, for example, Gaussian, distributions and

(b) to preserve the statistical difficulty of the original problem in the sense of Le Cam's asymptotic equivalence.

In this section, we propose such a paradigm based on discretizing the original Gaussian experiment, which achieves both of the above goals.

Discretized models. For any integer $t \in \mathbb{N}$, define the function $[\cdot]_t : \mathbb{R} \to 2^{-t}\mathbb{Z}$ by

$$[x]_t = 2^{-t}\lfloor 2^t x \rfloor. \tag{13}$$

The function $[\cdot]_t$ naturally extends to matrices componentwise: for $A = (A_{ij})$, $[A]_t = ([A_{ij}]_t)$.

Recall the submatrix detection problem (1). To model statistical inference with finite precision and complexity constraints, let us consider the same testing problem based on the discretized data $[X]_t$. In other words, the hypotheses are

$$H_0^t : [X]_t \sim \mathbb{P}_0^t \quad \text{versus} \quad H_1^t : [X]_t \sim \mathbb{P}_\theta^t, \qquad \theta \in \mathcal{M}(p, k, \lambda), \tag{14}$$

where for $X \sim \mathbb{P}_\theta$,

$$\mathbb{P}_\theta^t \triangleq \mathcal{L}([X]_t)$$

is the discrete distribution induced by the quantization operation (13), which is supported on $(2^{-t}\mathbb{Z})^{p\times p}$.

Now on the discretized data, any test for (14) is a (possibly randomized) function from the countable set $(2^{-t}\mathbb{Z})^{p\times p}$ to $\{0, 1\}$. Since there exists a one-to-one mapping between the set $(2^{-t}\mathbb{Z})^{p\times p}$ and the set of all finite length binary sequences $\bigcup_{n\in\mathbb{N}}\{0, 1\}^n$, each observed $[X]_t$ can be represented by a finite number of bits, and hence the computational complexity of any test of interest is well defined; see, for example, [7], Chapter 7. Thus the first goal of the paradigm is achieved. As an aside, we note that though each coordinate of the discretized data matrix $[X]_t$ has countably infinite support, its Shannon entropy is finite and behaves according to $H([X_{ij}]_t) = t + O(1)$ as $t \to \infty$ [36]. Therefore, if we choose $t = \Theta(\log p)$, then $[X]_t$ can be represented *on average* using variable-length lossless codes with $O(p^2 \log p)$ number of bits [16].

Given any test $\phi = \phi([X]_t)$ for (14), we can analogously define the worst-case Type-I + II error probability $\mathcal{E}(\phi)$ as in (3) but with $\mathbb{P}_0$ and $\mathbb{P}_\theta$ replaced by $\mathbb{P}_0^t$ and $\mathbb{P}_\theta^t$. Consequently, $\mathcal{E}^*$ can also be defined as in (4).

Asymptotic equivalence. Now we show that as long as we quantize each coordinate with accuracy p^{-c} for some constant $c > 0$, that is, $t = \Theta(\log p)$, the resulting family of discretized distributions $\{\mathbb{P}_\theta^t : \theta \in \mathbb{R}^{p \times p}\}$ is asymptotically equivalent to the original Gaussian experiment in the sense of Le Cam. Therefore any inference problem, in particular, submatrix detection, performed on the discretized data is asymptotically equally difficult as the original problem as $p \to \infty$, and hence we also achieve the second goal of the paradigm.

To state the equivalence result, recall the definition of Le Cam distance between statistical experiments. Let P be a probability measure on a standard Borel space $(\mathsf{X}, \mathcal{F})$, and let K denote a probability transition kernel (Markov kernel) from $(\mathsf{X}, \mathcal{F})$ to a standard Borel space $(\mathsf{Y}, \mathcal{G})$. Denote by KP the pushforward of P under K, that is, $KP(\mathrm{d}y) = \int_{\mathsf{X}} K(\mathrm{d}y|x)P(\mathrm{d}x)$. Given two experiments $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on $(\mathsf{X}, \mathcal{F})$ and $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$ on $(\mathsf{Y}, \mathcal{G})$ with common parameter space Θ, the *Le Cam deficiency* of $\mathcal{P}$ with respect to $\mathcal{Q}$ is defined by

$$\delta(\mathcal{P}, \mathcal{Q}) \triangleq \inf_T \sup_{\theta \in \Theta} \mathsf{TV}(TP_\theta, Q_\theta),$$

where the infimum is over all probability transition kernels from $(\mathsf{X}, \mathcal{F})$ to $(\mathsf{Y}, \mathcal{G})$ [39], Theorem 1.7, page 29. The *Le Cam distance* between $\mathcal{P}$ and $\mathcal{Q}$ is

$$\Delta(\mathcal{P}, \mathcal{Q}) \triangleq \delta(\mathcal{P}, \mathcal{Q}) \vee \delta(\mathcal{Q}, \mathcal{P}).$$

Two sequences of experiments $\{\mathcal{P}^{(p)}\}_{p \in \mathbb{N}}$ and $\{\mathcal{Q}^{(p)}\}_{p \in \mathbb{N}}$ are *asymptotically equivalent* if their Le Cam distance vanishes [33], Section 2.3, that is, if $\Delta(\mathcal{P}^{(p)}, \mathcal{Q}^{(p)}) \to 0$ as $p \to \infty$.

The following theorem, proved in the supplement [35], gives a nonasymptotic upper bound on the Le Cam distance between the Gaussian experiments $\mathcal{P}^{(p)} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}^{p \times p}\}$ and its discretized version $\mathcal{P}^{(p,t)} = \{\mathbb{P}_\theta^t : \theta \in \mathbb{R}^{p \times p}\}$. Therefore, as long as t grows at a logarithmical rate with p, the discretized model is asymptotically equivalent to the original Gaussian model.

THEOREM 1. *For any $t, p \in \mathbb{N}$, $\Delta(\mathcal{P}^{(p)}, \mathcal{P}^{(p,t)}) \le 2p^2 2^{-2t/3}$. Consequently, if $t = t(p) \ge (3 + \varepsilon)\log_2 p$ for any $\varepsilon > 0$, then $\{\mathcal{P}^{(p)}\}_{p \in \mathbb{N}}$ and $\{\mathcal{P}^{(p,t(p))}\}_{p \in \mathbb{N}}$ are asymptotically equivalent as $p \to \infty$.*

For the proof of Theorem 1, see the supplement [35]. An immediate consequence of Theorem 1 on submatrix detection is the following: Since the difference between the optimal Type-I + II error probabilities for the Gaussian hypotheses and the discretized hypotheses is upper bounded by their Le Cam distance

[33], Theorem 2.2, which vanishes as $p \to \infty$, we conclude that testing on discretized data has no impact on the statistical performance asymptotically in the high-dimensional setting. In particular, conclusion (6)–(7) continues to hold for testing (14).

REMARK 1. For the discretized model, when either condition in (6) holds, reliable detection can be attained by applying the linear or the scan test to the quantized data. To see this, note that the statistics T_{lin}, T_{scan} and $T_{\max}$ defined in (8)–(9) are all p-Lipschitz with respect to the entrywise ℓ_∞-norm of X. Using Lemma 1, it is straightforward to verify that if we compute T_{lin} and T_{scan} based on the quantized data $[X]_t$ with $t \geq (3+\varepsilon)\log p$, then $\mathcal{E}(\mathbf{1}_{\{T_{\mathrm{lin}}>\tau\}})$ [resp., $\mathcal{E}(\mathbf{1}_{\{T_{\mathrm{scan}}>\tau'\}})$] vanishes when the first (resp., second) condition in (6) holds. Here, the thresholds τ and τ' are defined in Lemma 1.

REMARK 2. From an alternative viewpoint, for appropriately chosen $t = t(p) \in \mathbb{N}$, one can restrict the attention to all tests that are measurable with respect to the σ-algebra on $\mathbb{R}^{p\times p}$ generated by $\mathcal{F}_t = \{\prod_{i,j=1}^{p}[x_{ij}2^{-t}, (x_{ij}+1)2^{-t}), x_{ij} \in \mathbb{Z}\}$ rather than the usual Borel σ-algebra generated by all open sets. Thus any such test ψ remains constant on any set in $\mathcal{F}_t$. Moreover, $\psi(X) = \psi([X]_t)$, and its computational complexity is well defined. Last but not least, the hypothesis testing problems (1) and (14) become equivalent on this smaller σ-algebra.

4. Complexity theoretic limits. In this section, we investigate complexity theoretic limits of the submatrix detection problem by drawing its connection to the planted clique problem. Let $N \in \mathbb{N}$ and $\kappa \in [N]$. We denote by $\mathcal{G}(N, 1/2)$ the Erdős–Rényi random graph on N vertices, where each edge is drawn independently at random with probability 1/2. In addition, following [3, 24], we use $\mathcal{G}(N, 1/2, \kappa)$ to denote the random graph generated by first sampling from $\mathcal{G}(N, 1/2)$, then picking κ vertices uniformly at random and connecting all edges in-between to form a clique of size κ. Distinguishing these two ensembles is known as the planted clique problem, formally defined as follows:

DEFINITION 1. Let $A \in \{0,1\}^{N\times N}$ be the adjacency matrix of a random graph drawn from either $\mathcal{G}(N, 1/2)$ or $\mathcal{G}(N, 1/2, \kappa)$. The *planted clique problem of parameters* (N, κ), denoted by $\mathsf{PC}(N, \kappa)$, refers to the hypothesis testing problem of

$$H_0^G : A \sim \mathcal{G}(N, 1/2) \quad \text{vs.} \quad H_1^G : A \sim \mathcal{G}(N, 1/2, \kappa). \tag{15}$$

The planted clique problem has a long history in the theoretical computer science literature. It is known that finding the clique is statistically impossible when $\kappa = o(\log N)$. Moreover, a greedy algorithm succeeds if $\kappa \geq c\sqrt{N \log N}$ for some constant $c > 0$ [32]. Using spectral methods, Alon, Krivelevich and Sudakov [3] provided the first polynomial time detection algorithm when $\kappa = c\sqrt{N}$, with later improvements obtained in, for example, [4, 18–21]. However, it is widely believed that the detection problem cannot be solved in randomized polynomial time when $\kappa = o(\sqrt{N})$, which can be summarized as the following *planted clique hypothesis*. This version is similar to [2], Conjecture 4.13, and [9], Hypothesis B_{PC}.

HYPOTHESIS 1. *For any sequence* $\{\kappa_N\}$ *such that* $\limsup_{N\to\infty} \frac{\log \kappa_N}{\log N} < 1/2$ *and any sequence of randomized polynomial-time tests*[3] $\{\psi_N\}$,

$$\liminf_{N\to\infty} \left(\mathbb{P}_{H_0^G}\{\psi_N(A) = 1\} + \mathbb{P}_{H_1^G}\{\psi_N(A) = 0\}\right) \geq \frac{2}{3}.$$

Various hardness results in theoretical computer science have been established based on the planted clique hypothesis, for example, approximating the Nash equilibrium [23], independence testing [2], certifying the restricted isometry property for compressed sensing measurement matrices [27], etc. Also, several cryptographic schemes have been proposed assuming the intractability of finding planted cliques [25, 31] or bicliques [5]. Recently, the average-case hardness of planted clique has been established under certain computation models; see, for example, [22, 37].

The main result of the current paper is the following.

THEOREM 2. *Assume that Hypothesis* 1 *holds. Consider testing the discrete hypotheses* (14) *with* $t = t(p) = 4\lceil \log_2 p \rceil$ *in the asymptotic regime* (5). *If, for some absolute constant* $\delta > 0$,

$$\frac{\lambda}{p/k^{2+\delta}} \to 0 \quad \textit{and} \quad \limsup_{p\to\infty} \lambda\sqrt{\log p} \leq \frac{1}{6}, \tag{16}$$

there exists no *sequence of randomized polynomial-time tests* $\{\phi_p\}$ *such that* $\liminf_{p\to\infty} \mathcal{E}(\phi_p) < 2/3$ *for testing* (14). *Conversely, if*

$$\frac{\lambda}{p/k^2} \to \infty \quad \textit{or} \quad \liminf_{p\to\infty} \frac{\lambda}{2\sqrt{\log p}} > 1, \tag{17}$$

there is a sequence of linear-time tests $\{\phi_p\}$ *such that* $\mathcal{E}(\phi_p) \to 0$.

As shown later in the proof of Theorem 2, one can use $T_{\text{lin}}([X]_t)$ (resp., $T_{\max}([X]_t)$) as the test statistic when the first (resp., second) condition in (17) holds. It is straightforward to see that both T_{lin} and $T_{\max}$ are of linear complexity.

[3] Formally, randomized polynomial-time tests belong to the **BPP** complexity class. Interested readers are referred to standard textbooks on computational complexity theory (e.g., [7], Chapter 7) for the formal definitions and discussions. Intuitively speaking, randomized polynomial-time tests refer to algorithms with output space {0, 1}, which have access to external random numbers and whose running time is bounded by a polynomial of the input length regardless of the random numbers.

Contrasting the statistical limit (6)–(7) with the computational limit (16)–(17), we obtain the following implication of Theorem 2 on the complexity of submatrix detection: suppose that $k \le p^{\alpha}$ for some absolute constant $\alpha \in (0, 2/3)$. Then $\lambda \asymp \sqrt{\frac{1}{k}\log\frac{p}{k}}$ implies $\frac{\lambda}{p/k^2} \to 0$. Consequently, conditions (6)–(7) and Theorem 1 imply that reliable detection is statistically possible if and only if $\lambda = \Omega(\sqrt{\frac{1}{k}\log\frac{p}{k}})$. In contrast, condition (16) in Theorem 2 asserts that, to accomplish the same task using randomized polynomial-time algorithms, it is necessary to have $\lambda = \Omega(\frac{1}{\sqrt{\log p}} \wedge \frac{p}{k^{2+\delta}})$ for all $\delta > 0$, which far exceeds $\sqrt{\frac{1}{k}\log\frac{p}{k}}$ whenever $k \gg (\log p)^2$. Therefore, computationally efficient procedures require significantly larger signal level λ to reliably detect the submatrix than the statistical optimum. More precisely, if $k = \Theta(p^{\alpha})$ for some $\alpha \in (0, 2/3)$, then the minimal λ for any randomized polynomial-time test to succeed is at least $\lambda = \Omega(\frac{1}{\sqrt{\log p}})$ when $\alpha \in (0, 1/2)$ and $\Omega(p^{1-2\alpha-\delta})$ for any $\delta > 0$ when $\alpha \in [1/2, 2/3)$, which exceeds the statistical optimal level $\lambda = \Theta(p^{-\alpha/2}\sqrt{\log p})$ by a polynomial factor in p. Thus, in this regime, computational complexity constraints severely limit the best possible statistical performance in the submatrix detection problem. On the other hand, when $k \ge p^{\alpha}$ for some $\alpha > 2/3$, $\frac{\lambda}{p/k^2} \to \infty$ is the dominating condition in both (6) and (17), and a computationally efficient test based on T_{lin} achieves statistically optimal detection in this regime. Figure 1 in Section 1 provides a graphical illustration of the above discussion.

It should be noted that the sub-polynomial factor difference, that is, $p/k^{2+\delta}$ versus p/k^2, in the first part of (16) and (17) is a direct consequence of Hypothesis 1. In contrast, the logarithmic factor difference in the second part of (16) and (17) can potentially be closed by employing better reduction argument and/or more sophisticated testing procedures such as those based on spectral methods, which we leave as a future direction.

The remainder of this section is devoted to proving Theorem 2, with auxiliary lemmas proved in Section 6. First, in Section 4.1 we provide some intuition on how the planted clique problem is related to the submatrix detection problem (1) under the Gaussian model. Next, in Section 4.2 we prove that under the asymptotically equivalent discretized model, every randomized polynomial time submatrix detector for (14) leads to a randomized polynomial time solver for the planted clique problem of appropriate parameters with almost identical performance. Finally, a proof of Theorem 2 is presented in Section 4.3.

4.1. *Planted clique and submatrix detection.* We first explain how the submatrix detection problem can be reduced from the planted clique problem under the original Gaussian model. These results are presented as the precursor of the randomized polynomial time reduction for the discretized model in Section 4.2, as well as to provide insights into the hardness of the submatrix detection problem. A connection between the two problems has also been previously hinted at in [1].

Recall the Gaussian submatrix detection problem in (1) with parameter (p, k, λ). For some $\ell \in \mathbb{N}$ to be chosen depending on p, k and λ, let

$$N = 2p\ell. \tag{18}$$

We construct a reduction scheme which maps any adjacency matrix $A \in \{0, 1\}^{N\times N}$ to a random matrix $X \in \mathbb{R}^{p\times p}$ in $O(N^2)$ number of flops, such that the following holds: if A is drawn from $\mathcal{G}(N, 1/2)$ under H_0^G, then the distribution of X is close in total variation distance to the null distribution $\mathbb{P}_0$; if A is drawn from $\mathcal{G}(N, 1/2, \kappa)$ under H_1^G, then the law of X is close in total variation distance to a mixture of distributions in the alternative H_1, where the clique size κ is a constant multiple of k.

Randomized reduction. An important step in the following reduction scheme is to map any random edge to an $N(0, 1)$ random variable and any edge in the clique to an $N(\mu, 1)$ random variable with some positive mean value μ. Although this goal might not be achievable exactly, we describe below a strategy to achieve it approximately.

To this end, for any $M \geq 3$ and $0 < \mu \leq \frac{1}{2M}$, let $c_0 = (1 - 2\overline{\Phi}(M))^{-1}$ and $c_1 = [1 - \overline{\Phi}(M - \mu) - \overline{\Phi}(M + \mu)]^{-1}$. We define two distributions $\mathcal{F}_1$ and $\mathcal{F}_0$ with the respective density functions

$$\begin{aligned} f_1(x) &= c_1\varphi(x - \mu)\mathbf{1}_{\{|x|\leq M\}}, \\ f_0(x) &= \bigl[2c_0\varphi(x) - c_1\varphi(x - \mu)\bigr]\mathbf{1}_{\{|x|\leq M\}}. \end{aligned} \tag{19}$$

Here both f_0 and f_1 are well-defined probability density functions. In particular, the conditions $M \geq 3$ and $0 \leq \mu \leq \frac{1}{2M}$ ensure that $f_0 \geq 0$. In what follows, both M and μ, and thereby $\mathcal{F}_0$ and $\mathcal{F}_1$, depend on N, though we suppress the dependence for notational convenience.

The randomized mapping from A to X is as follows. By (18), N is even, and let $N_2 = N/2 = p\ell$ and $[N] \setminus [N_2] = \{N_2 + 1, \ldots, N\}$:

(1) (Gaussianization). Let $A_0 = A_{[N]\setminus[N_2],[N_2]} \in \mathbb{R}^{N_2\times N_2}$ be the lower-left quarter of the matrix A. Independent of A, we generate two $N_2 \times N_2$ matrices B_0 and B_1, whose entries are sampled independently from $\mathcal{F}_0$ and $\mathcal{F}_1$ with density functions given in (19), respectively. Define an $N_2 \times N_2$ matrix B by

$$B_{ij} = (B_0)_{ij}\bigl(1 - (A_0)_{ij}\bigr) + (B_1)_{ij}(A_0)_{ij}. \tag{20}$$

In other words, $\mathcal{L}(B_{ij}|(A_0)_{ij} = 0) = \mathcal{F}_0$ and $\mathcal{L}(B_{ij}|(A_0)_{ij} = 1) = \mathcal{F}_1$.

(2) (Partitioning). Partition B into ℓ^2 consecutive $p \times p$ blocks. In other words, for $i, j \in [\ell]$, the (i, j)th block is $B^{(i,j)} = (B_{a,b}^{(i,j)}) \in \mathbb{R}^{p\times p}$ where

$$B_{a,b}^{(i,j)} = B_{(i-1)p+a,(j-1)p+b} \qquad \forall a, b \in [p]. \tag{21}$$

(3) (Averaging). Define $X \in \mathbb{R}^{p\times p}$ by summing up all ℓ^2 blocks and scaling by ℓ:

$$X = \frac{1}{\ell}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} B^{(i,j)}. \tag{22}$$

Therefore, (20), (21) and (22) collectively define a deterministic function

$$\begin{aligned} g : \{0,1\}^{N\times N} \times \mathbb{R}^{N_2\times N_2} \times \mathbb{R}^{N_2\times N_2} &\to \mathbb{R}^{p\times p}, \\ (A, B_0, B_1) &\mapsto X \end{aligned} \tag{23}$$

which can be computed in $O(N^2)$ number of flops. The reason that we call the first step "Gaussianization" is due to the following lemma, which ensures that for appropriately chosen M and μ, the marginal distribution of B_{ij} is close to the Gaussian distribution of unit variance and mean zero (resp., μ) if $(A_0)_{ij}$ corresponds to a random edge (resp., an edge in the clique).

LEMMA 2. *Let $N \geq 6$. Let ξ be a Bernoulli random variable. Let W be a random variable such that for $i \in \{0,1\}$, the conditional distribution of $W|\xi = i$ follows f_i in (19) where $M \geq 3$ and $\mu \leq \frac{1}{2M}$. Then*:

(1) *if* $\mathbb{P}\{\xi = 1\} = 1$, *then* $\mathsf{TV}(\mathcal{L}(W), N(\mu,1)) \leq \mathrm{e}^{(1-M^2)/2}$;
(2) *if* $\xi \sim \text{Bernoulli}(1/2)$, *then* $\mathsf{TV}(\mathcal{L}(W), N(0,1)) \leq \mathrm{e}^{-M^2/2}$.

The following two lemmas characterize the law of $X = g(A, B_0, B_1)$ when either H_0^G or H_1^G in (15) holds.

LEMMA 3. *Suppose H_0^G holds and $N \geq 2p \geq 6$. Let $M \geq \sqrt{6\log N}$. Then*

$$\mathsf{TV}\big(\mathcal{L}(X), \mathbb{P}_0\big) \leq \frac{1}{p}. \tag{24}$$

LEMMA 4. *Suppose H_1^G holds with $N \geq 2p$, $p \geq 2\kappa$ and $\kappa \geq 20$. Let $k = \lfloor \kappa/20 \rfloor$. Let $M \geq \sqrt{6\log N}$ and $\mu \leq \frac{1}{2M}$ in (19). Then there exists a prior π on $\mathcal{M} = \mathcal{M}(p, k, \frac{2\mu p}{N})$ such that for $\mathbb{P}_\pi(\cdot) = \int_{\mathcal{M}} \mathbb{P}_\theta(\cdot)\pi(\mathrm{d}\theta)$,*

$$\mathsf{TV}\big(\mathcal{L}(X), \mathbb{P}_\pi\big) \leq \frac{1}{p} + 40k\left(\frac{\mathrm{e}}{4}\right)^{5k} + 2k\exp\left(-4k\log\frac{p}{20k}\right). \tag{25}$$

REMARK 3. A careful examination of the proof of Lemma 4 in Section 6.4 reveals that the prior π is in fact supported on a subset $\widetilde{\mathcal{M}}(p,k,\lambda) \subset \mathcal{M}(p,k,\lambda)$ where

$$\begin{aligned} \widetilde{\mathcal{M}}(p,k,\lambda) \triangleq \big\{&\theta \in \mathbb{R}^{p\times p} : \exists U, V \subset [p], \text{ s.t. } k \leq |U|, |V| \leq 20k, \\ &\theta_{ij} \geq \lambda, \text{ if } (i,j) \in U \times V, \theta_{ij} = 0 \text{ if } (i,j) \notin U \times V\big\}, \end{aligned} \tag{26}$$

and $\lambda = \frac{2\mu p}{N}$. In other words, any matrix in $\widetilde{\mathcal{M}}(p, k, \lambda)$ contains a nonzero rectangular submatrix whose row and column support sizes are between k and $20k$. This observation will be useful for studying the hardness of estimation in Section 5.2.

Combining Lemmas 3 and 4, the following theorem shows that any submatrix detector leads to a test with almost identical error probability for a planted clique problem, whose parameters (N, κ) depend on the parameters (p, k, λ) of the submatrix detection problem.

THEOREM 3. *Assume that* $p \geq 40k$ *and* $\lambda \leq \frac{1}{2\sqrt{6\log(2p)}}$. *Suppose* $\phi : \mathbb{R}^{p\times p} \to \{0, 1\}$ *is a test for distinguishing* H_0 *and* H_1 *in* (1) *with Type-*I + II *error probability upper bounded by* ε, *that is,*

$$\mathbb{P}_{\theta_0}\{\phi(X) = 1\} + \sup_{\theta\in\mathcal{M}(p,k,\lambda)} \mathbb{P}_\theta\{\phi(X) = 0\} \leq \varepsilon. \tag{27}$$

Let $\kappa = 20k$, $N = 2p\ell$ *and* $N_2 = N/2$, *where* ℓ *is the largest positive integer such that* $N\sqrt{6\log N} \leq p/\lambda$. *Let* $B_0, B_1 \in \mathbb{R}^{N_2\times N_2}$ *have i.i.d. entries drawn from* $\mathcal{F}_0$ *and* $\mathcal{F}_1$, *respectively, with* $M = \sqrt{6\log N}$ *and* $\mu = \frac{1}{2M}$. *Then* $\psi(\cdot) = \phi(g(\cdot, B_0, B_1))$ *is a test for the planted clique detection problem* (15) *whose Type-*I + II *error probability is upper bounded by*

$$\mathbb{P}_{H_0^G}\{\psi(A) = 1\} + \mathbb{P}_{H_1^G}\{\psi(A) = 0\} \leq \varepsilon + \beta, \tag{28}$$

where $\beta = \frac{2}{p} + 40k(\frac{e}{4})^{5k} + 2k\exp(-4k\log\frac{p}{20k})$.

PROOF. Let A denote the adjacency matrix of $\mathcal{G}$. By definition, we have $M = \sqrt{6\log N}$, while the definition of ℓ ensures that $\frac{2\mu p}{N} \geq \lambda$, and the constraint $\lambda \leq \frac{1}{2\sqrt{6\log(2p)}}$ guarantees $\ell \geq 1$.

By the definition of the total variation distance, Lemma 3 implies that under H_0^G,

$$\begin{aligned} &\left|\mathbb{P}_{H_0^G}\{\phi(g(A, B_0, B_1)) = 1\} - \mathbb{P}_0\{\phi(X) = 1\}\right| \\ &\quad\leq \mathsf{TV}(\mathcal{L}(g(A, B_0, B_1)), \mathbb{P}_0) \leq \frac{1}{p} \triangleq \beta_0. \end{aligned} \tag{29}$$

On the other hand, since $\frac{2\mu p}{N} \geq \lambda$, we have $\mathcal{M}(p, k, \frac{2\mu p}{N}) \subset \mathcal{M}(p, k, \lambda)$. So any location mixture $\mathbb{P}_\pi$ of the former can be viewed as a mixture of the latter. Hence, Lemma 4 implies that under H_1^G,

$$\begin{aligned} &\left|\mathbb{P}_{H_1^G}\{\phi(g(A, B_0, B_1)) = 0\} - \mathbb{P}_\pi\{\phi(X) = 0\}\right| \\ &\quad\leq \mathsf{TV}(\mathcal{L}(g(A, B_0, B_1)), \mathbb{P}_\pi) \\ &\quad\leq \frac{1}{p} + 40k\left(\frac{e}{4}\right)^{5k} + 2k\exp\left(-4k\log\frac{p}{20k}\right) \triangleq \beta_1. \end{aligned} \tag{30}$$

Since $\beta = \beta_0 + \beta_1$, the desired error bound (28) follows from

$$\begin{aligned}
&\mathbb{P}_{H_0^G}\{\phi(g(A, A_0, W)) = 1\} + \mathbb{P}_{H_1^G}\{\phi(g(A, A_0, W)) = 0\} \\
&\quad \le \mathbb{P}_{\theta_0}\{\phi(X) = 1\} + \mathbb{P}_{\pi}\{\phi(X) = 1\} + \beta_0 + \beta_1 \\
&\quad \le \mathbb{P}_{\theta_0}\{\phi(X) = 1\} + \sup_{\theta \in \mathcal{M}(p,k,\lambda)} \mathbb{P}_{\theta}\{\phi(X) = 1\} + \beta \\
&\quad \le \varepsilon + \beta,
\end{aligned}$$

where the last inequality is due to assumption (27) on ϕ. □

REMARK 4. Although the reduction scheme g can be implemented in $O(N^2)$ flops, its computational complexity is ill defined since it involves computing sums of continuous random variables and processing infinitely many bits. This issue will be addressed by a quantization argument in the next subsection when we deal with the discretized models.

4.2. *Randomized polynomial-time reduction for discretized models.* In this section, we show that with slight modifications, the scheme introduced in Section 4.1 can be made into a randomized polynomial-time reduction from the planted clique problem to the submatrix detection problem under discretized models in rigorous complexity-theoretic sense.

For the discretized model $\mathcal{P}^{(p,t)}$ introduced in Section 3, the reduction scheme from the planted clique model follows the same steps in Section 4.1, except that both the input (B_0, B_1) and the output X are now discretized.

To this end, we first define discrete approximations, denoted by Q_0 and Q_1, to the densities f_0 and f_1 defined in (19). Let w, T be integers to be chosen based on t, M and N. Recall the quantization operator defined in (13) and that B_0 and B_1 consist of i.i.d. entries drawn from densities f_0 and f_1, respectively, which are supported on $[-M, M]$ by definition. Note that each $[(B_0)_{ij}]_w$ is drawn from a distribution with atoms x_i and probability mass function (p.m.f.) p_i for $i \in [M2^{w+1}]$. To find a dyadic approximation for the p.m.f., let $q_i = \lfloor p_i 2^T \rfloor 2^{-T}$ for $i = 2, \dots, M2^w$ and $q_1 = \lfloor p_1 2^T \rfloor 2^{-T} + 1 - \sum_{i \ge 2} q_i$, where

$$T = \lceil \log_2 M \rceil + w + 3 \log_2 N. \tag{31}$$

Denote by Q_0 the discrete distribution with atoms x_i and probability masses q_i. Similarly, define Q_1 as the dyadic approximation for the distribution of $[(B_1)_{ij}]_w$.

The reduction scheme operates as follows: first, generate $\breve{B}_i$ consisting of i.i.d. entries drawn from Q_i for $i = 0, 1$. Next, replace the matrices B_0 and B_1 in (20) by their discretized version $\breve{B}_0$ and $\breve{B}_1$, and denote the resulting matrix by $\breve{B}$. Applying (21)–(22) to $\breve{B}$, we obtain $\breve{X}$ and output its quantized version $[\breve{X}]_t$. Implementing the above steps yields a deterministic function

$$\begin{aligned}
\breve{g} : \{0, 1\}^{N \times N} \times \left([-M, M]_w\right)^{N_2 \times N_2} \times \left([-M, M]_w\right)^{N_2 \times N_2} &\to (2^{-t}\mathbb{Z})^{p \times p}, \\
(A, \breve{B}_0, \breve{B}_1) &\mapsto [\breve{X}]_t,
\end{aligned} \tag{32}$$

where $[-M, M]_w = [-M, M] \cap 2^{-w}\mathbb{Z}$ is the quantized interval.

REMARK 5 (Computational complexity of reduction). First we discuss the complexity for generating the auxiliary random variables used in the reduction scheme. Note that each $(\breve{B}_0)_{ij}$ is drawn from Q_0 whose atoms x_i can be represented by $\lceil \log_2 M \rceil + w$ bits and the p.m.f. q_i is a dyadic rational with T bits. Therefore sampling from the distribution Q_0 can be done using the inverse CDF[4] by outputting x_J, where $J = \min\{j : \sum_{i=1}^{j} q_i \le U2^{-T}\}$ and U is a random integer uniformly distributed on $[2^T]$. Consequently, sampling from Q_0 requires $O(M2^w T)$ preprocessing time to compute the CDF, and T fair coin flips and $O(\log M + w)$ time per sample (via binary search). Furthermore, discretizing each entry $\breve{X}_{ij}$ to $[\breve{X}_{ij}]_t$ involves keeping the first t bits after the binary point, which can be computed in $O(t)$ time. Therefore we conclude that $\breve{g}$ can be computed using $O((\lceil \log_2 M \rceil + w + t)N^2)$ number of binary operations.

To summarize, the randomized reduction scheme requires $O(N^2 T)$ random bits and $O(M2^w T + N^2(\log M + w + t))$ computation, where T is defined in (31). As we will show in Section 4.3, for all cases of interest in this paper, we can set $N = O(p^2)$ and $M, w, t = O(\log_2 p) = O(\log_2 N)$. Therefore, our reduction for discretized models $A \mapsto \breve{g}(A, \breve{B}_0, \breve{B}_1)$ is a *randomized polynomial-time reduction*.

We now investigate the distributions of $[\breve{X}]_t$ under H_0^G and H_1^G, respectively. The following lemmas are counterparts of Lemmas 3 and 4 for discretized models. Comparing with the total variation bound (24) and (25), we show that, upon suitable choices of w depending on (t, N), replacing B_0 and B_1 with the discrete versions $\breve{B}_0$ and $\breve{B}_1$ only introduces an extra term of $4/p$ in the total variation of $\mathcal{L}([\breve{X}]_t)$ to $\mathbb{P}_0^t$ under H_0^G, and to a mixture of the alternative distributions $\mathbb{P}_\theta^t$ under H_1^G, respectively. This objective is accomplished by putting the support of Q_0 and Q_1 on a finer grid than that of the output $\breve{X}$, that is, choosing $w > t$, which is essential for controlling the approximation error in the output distribution incurred by quantizing the input.

LEMMA 5. *Let $N \ge 2p \ge 6$. Let $w \in \mathbb{N}$ satisfy*

$$w \ge t + 6\log_2 N. \tag{33}$$

Then under H_0^G,

$$\mathsf{TV}(\mathcal{L}([\breve{X}]_t), \mathbb{P}_0^t) \le \frac{5}{p}. \tag{34}$$

[4] More sophisticated random number generators for discrete distributions (such as Walker's alias method which requires linear time for preprocessing and constant time per sample) can be found in [26], Section 3.4.1.

LEMMA 6. *Suppose H_1^G holds with $N \geq 2p$, $p \geq 2\kappa$, $\kappa \geq 20$, $k = \lfloor \kappa/20 \rfloor$. Let $M \geq \sqrt{6 \log N}$ and $\mu \leq \frac{1}{2M}$. Let w satisfy* (33). *Then there exists a prior π on $\mathcal{M}(p, k, \frac{2\mu p}{N})$, such that for $\mathbb{P}_\pi^t(\cdot) = \int_{\mathcal{M}} \mathbb{P}_\theta^t(\cdot)\pi(\mathrm{d}\theta)$,*

$$\mathsf{TV}(\mathcal{L}([\breve{X}]_t), \mathbb{P}_\pi^t) \leq \frac{5}{p} + 40k\left(\frac{\mathrm{e}}{4}\right)^{5k} + 2k \exp\left(-4k \log \frac{p}{20k}\right). \tag{35}$$

Combining the two lemmas, we obtain the following result analogously to Theorem 3.

THEOREM 4. *Assume that $p \geq 40k$ and $\lambda \leq \frac{1}{2\sqrt{6\log(2p)}}$. Suppose ϕ: $(2^{-t}\mathbb{Z})^{p \times p} \to \{0, 1\}$ is a test for distinguishing H_0^t and H_1^t in* (14) *with Type-*I + II *error probability upper bounded by ε, that is,*

$$\mathbb{P}_0^t\{\phi([X]_t) = 1\} + \sup_{\theta \in \mathcal{M}(p,k,\lambda)} \mathbb{P}_\theta^t\{\phi([X]_t) = 0\} \leq \varepsilon. \tag{36}$$

Let $\kappa, N, N_2 \in \mathbb{N}$ be chosen as in Theorem 3. *Let $w \in \mathbb{N}$ satisfy* (33) *and $\breve{g}: \{0,1\}^{N \times N} \times ([-M, M]_w)^{N_2 \times N_2} \times ([-M, M]_w)^{N_2 \times N_2} \to (2^{-t}\mathbb{Z})^{p \times p}$ be defined in* (32). *Then $\psi(\cdot) = \phi(\breve{g}(\cdot, \breve{B}_0, \breve{B}_1))$ is a test for the planted clique detection problem* (15) *whose Type-*I + II *error probability is upper bounded by*

$$\mathbb{P}_{H_0^G}\{\psi(A) = 1\} + \mathbb{P}_{H_1^G}\{\psi(A) = 0\} \leq \varepsilon + \beta, \tag{37}$$

where $\beta = \frac{10}{p} + 40k(\frac{\mathrm{e}}{4})^{5k} + 2k \exp(-4k \log \frac{p}{20k})$.

PROOF. In view of the analogy between Lemmas 3–4 and 5–6, the proof follows the same argument as that in the proof of Theorem 3, except that $\mathbb{P}_0$ and $\mathbb{P}_\pi$ are replaced by $\mathbb{P}_0^t$ and $\mathbb{P}_\pi^t$, respectively, $g(A, B_0, B_1)$ is replaced by $\breve{g}(A, \breve{B}_0, \breve{B}_1)$, β_0 and β_1 in (29) and (30) are both increased by $4/p$. □

4.3. *Proof of Theorem* 2. We start with the lower bound. Without loss of generality, we can assume that $\lambda \geq 1/p$ since when $\lambda < 1/p$, both conditions in (7) hold in the asymptotic regime (5), and the problem is statistically impossible. Let the sequence $\{(k(p), \lambda(p))\}$ satisfy (5) and (16). Let $\{\phi_p\}$ be a sequence of randomized polynomial time tests. For conciseness we drop the indices in $k(p)$, $\lambda(p)$ and ϕ_p. Suppose for the sake of contradiction that

$$\liminf_{p \to \infty}\left(\mathbb{P}_0^t\{\phi([X]_t) = 1\} + \sup_{\theta \in \mathcal{M}(p,k,\lambda)} \mathbb{P}_\theta^t\{\phi([X]_t) = 0\}\right) < \frac{2}{3}. \tag{38}$$

Choose κ and N as in Theorems 3–4, that is, $\kappa = 20k$ and $N = 2p\ell$ where ℓ is the largest integer such that $N\sqrt{6 \log N} \leq p/\lambda$. Since the first condition in (16) implies that $\lambda \leq Cp/k^{2+\delta}$ for some constant C and all sufficiently large p, we have $2\kappa^{2+\delta/2}\sqrt{6\log(2\kappa^{2+\delta/2})} \le p/\lambda$ and hence $\ell \ge \lfloor \kappa^{2+\delta/2}/p \rfloor$ for all sufficiently large p. Similarly, the second condition in (16) implies that $\lambda \le \frac{1}{2\sqrt{6\log(2p)}}$ for all sufficiently large p and consequently, $\ell \ge 1$. Using the simple fact that $x\lfloor y/x \rfloor \vee x \ge y/2$ for all $x, y > 0$, we conclude that $N = 2p\ell \ge \kappa^{2+\delta/2} \vee (2p)$ for all sufficiently large p, hence

$$\liminf_{p\to\infty} \frac{\log\kappa}{\log N} \le \frac{1}{2+\delta/2} < \frac{1}{2}. \tag{39}$$

On the other hand, we have $N \le p/\lambda \le p^2$, where the last inequality holds since we have assumed $\lambda \ge 1/p$. Applying Theorem 4 with $w = 16\lceil \log_2 p \rceil \ge t + 12\log_2 p \ge t + 6\log_2 N$, we conclude from (38) that the randomized test $\psi(\cdot) = \phi(\breve{g}(\cdot, \breve{B}_0, \breve{B}_1))$ satisfies

$$\liminf_{p\to\infty} \bigl(\mathbb{P}_{H_0^G}\{\psi(A) = 1\} + \mathbb{P}_{H_1^G}\{\psi(A) = 0\}\bigr) < \frac{2}{3}. \tag{40}$$

In view of Remark 5, $A \mapsto \breve{g}(A, \breve{B}_0, \breve{B}_1)$ is a randomized polynomial-time reduction. By the assumption on ϕ, ψ as a composition of ϕ and $\breve{g}$ is a randomized polynomial-time test for $\mathsf{PC}(N, \kappa)$. Therefore, (40) contradicts Hypothesis 1 in view of (39).

It remains to show the upper bound. Denote the linear and maximum test statistics computed on the discretized matrix $[X]_t$ by T_{lin} and $T_{\max}$, respectively. If the first condition in (17) holds, that is, $\lambda k^2/p \to \infty$, in view of Lemma 1 and Remark 1, we have $\mathcal{E}(\mathbf{1}_{\{T_{\text{lin}} > \tau\}}) \to 0$ where τ is defined in Lemma 1. If the second condition in (17) holds, recall $\tau'' = \sqrt{(4+c)\log p}$ defined in Lemma 1. If the constant c is sufficiently small such that $\liminf_{p\to\infty} \lambda/\tau'' > 1$, then following the reasoning in Remark 1, it is straightforward to verify that $\mathcal{E}(\mathbf{1}_{\{T_{\max} > \tau''\}}) \to 0$. This completes the proof.

5. Discussion.

In this paper, assuming the planted clique hypothesis, we have demonstrated a phase transition phenomenon on gaps between the optimal statistical performance with and without computational complexity constraints for the submatrix detection problem. The hardness result in Theorem 2 has important consequences on the hardness of two related problems, namely, *support recovery* and *matrix estimation* under submatrix sparsity, both of which are more difficult than detection and require stronger signal level. To discuss computational complexity of statistical procedures, we focus on the discretized models introduced in Section 3 throughout the current section.

5.1. *Support recovery.*

As previously studied in [28], the goal of support recovery is to identify the minimum λ such that, under the alternative hypothesis H_1 in (1), the submatrix can be consistently located. According to Theorems 1 and 2 of [28], for all $k \le p/2$, one needs $\lambda = \Omega(\sqrt{\log(p)/k})$ to recover the support consistently under the parameter space (2). Compared with (6)–(7), this coincides with the minimum signal strength required for detecting the submatrix when $k = O(p^{\alpha})$ for $\alpha < 2/3$, but is much larger when $k = \Omega(p^{\alpha})$ for $\alpha > 2/3$.

Intuitively, locating the submatrix is more difficult than detecting the mere existence thereof. Therefore, the complexity theoretic limit for support recovery should also exceed that of detection. This claim, however, does not follow immediately since support recovery only deals with the alternative hypothesis (2), and the null hypothesis is excluded from the parameter set. To provide a rigorous argument, for $\lambda = \Omega(\sqrt{\log(p)/k})$, given a support estimator $(\widehat{U}, \widehat{V})$ such that $\sup_{\theta \in \mathcal{M}} \mathbb{P}_\theta\{(\widehat{U}, \widehat{V}) \neq (U, V)\} \le \varepsilon$, we can construct a test for (1) which rejects if $T = \sum_{i \in \widehat{U}, j \in \widehat{V}} X_{ij}$ exceeds τ' defined in Lemma 1. Since T is at most T_{scan}, the same argument in the proof of Lemma 1 shows that the Type-I + II error probability for this test is upper bounded by ε plus the right-hand side of (11), which vanishes when $p, k \to \infty$. This implies that the minimal λ achievable by computationally efficient support estimators is at least a constant factor of that required by computationally efficient submatrix detectors. Therefore, Theorem 2 implies that no randomized polynomial-time algorithm can achieve consistent support recovery when condition (16) holds. This resolves in the negative the open question raised in [8], Section 5, on the existence of computationally efficient minimax procedures in the regime of $k = O(p^{\alpha})$ for any $\alpha < 2/3$. It remains open to determine whether the statistically optimal support recovery can be achieved computationally efficiently when $k = \Theta(p^{\alpha})$ for $\alpha > 2/3$.

5.2. *Hardness of estimation depends on norm.* We now consider the computational aspect of the related problem of estimating the mean matrix with submatrix sparsity under squared norm losses. Denote the set of $k \times k$-sparse matrices by

$$\begin{aligned} \mathcal{F}(p,k) = \bigl\{\theta \in \mathbb{R}^{p \times p} : \exists U, V \subset [p],\ \text{s.t. } |U|, |V| \le k, \\ \theta_{ij} = 0 \text{ if } (i,j) \notin U \times V\bigr\}, \end{aligned} \tag{41}$$

which includes both the zero matrix and the set $\widetilde{\mathcal{M}}(p, \lfloor k/20 \rfloor, \lambda)$ [defined in (26)] for any $\lambda > 0$.

Given the noisy observation $X = \theta + Z$, where Z consists of standard Gaussian entries, the minimax risk

$$\Psi_{\|\cdot\|}(p,k) \triangleq \inf_{\tilde{\theta}} \sup_{\theta \in \mathcal{F}(p,k)} \mathbb{E}\|\tilde{\theta} - \theta\|^2 \tag{42}$$

has been obtained in [34], Section 4, within universal constant factors using convex geometry and information-theoretic arguments for all unitarily invariant norms,[5]

[5] To be precise, note that the minimax rates in [34] are obtained for the Gaussian model. Since the loss function is unbounded, one cannot directly conclude from asymptotic equivalence that the

in particular, satisfies

$$k\|I_k\|^2 \lesssim \Psi_{\|\cdot\|}(p,k) \lesssim k\|I_k\|^2 \log \frac{\mathrm{e}p}{k}.$$

Capitalizing on the hardness result of detecting submatrices in Theorem 2, we show that the minimax estimation rates corresponding to certain norm losses cannot be attained by computationally efficient methods. For conciseness, let us focus on the class of Schatten-q norms $\|\cdot\|_{\mathrm{S}_q}$, defined as the ℓ_q-norm of singular values for $q \in [1,\infty]$. The minimax rate is given by (see [34], Example 1)

$$\Psi_{\mathrm{S}_q}(p,k) \asymp k^{2/q+1} + k^{(2/q)\vee 1} \log \frac{\mathrm{e}p}{k}, \tag{43}$$

which is within a logarithmic term of $k^{2/q+1}$. Next we discuss the computational cost of estimation by focusing on the asymptotic regime where $k = \Theta(p^\alpha)$ for some $\alpha \in (0,1)$. In view of the relationship between testing and estimation, we can use the construction in Lemmas 3–4 for the detection problem

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \in \widetilde{\mathcal{M}}(p, \lfloor k/20 \rfloor, \lambda)$$

as a two-point lower bound. Note that for any $\theta \in \widetilde{\mathcal{M}}(p, \lfloor k/20 \rfloor, \lambda)$ and any $q \in [1,\infty]$, $\|\theta\|_{\mathrm{S}_q} \ge \|\theta\|_{\mathrm{S}_\infty} = \Omega(k\lambda)$. Assuming Hypothesis 1, we conclude that the squared Schatten-q norm risk achievable by any randomized polynomial-time estimator is at least $\Omega(k^2\lambda^2)$ for any λ satisfying (16). Thus, for any constant $\delta > 0$, the worst-case risk is at least $\Omega(k^{-\delta}(k^2 \wedge \frac{p^2}{k^2}))$. Note that this lower bound is not monotonic in k and can be easily improved to

$$\Omega\bigl(k^{-\delta}\bigl(k^2 \wedge p\bigr)\bigr) \tag{44}$$

since the risk is clearly nondecreasing in k.

On the constructive side, an estimation error of

$$O\biggl(k^{(2/q+1)\vee 2} \log \frac{\mathrm{e}p}{k} \wedge p^{2/q+1}\biggr) \tag{45}$$

in squared Schatten-q norm can be achieved in polynomial time. To see this, first note that treating a $k \times k$-sparse matrix as a k^2-sparse vector in p^2-dimensional space and applying entrywise hard thresholding yields an estimator $\hat{\theta}$ whose mean-square error (i.e., squared Frobenius or Schatten-2 norm) is at most $O(k^2 \log \frac{\mathrm{e}p}{k})$. Then we project $\hat{\theta}$ into the space of row-sparse matrices to obtain $\tilde{\theta}$ by choosing the k rows of $\hat{\theta}$ of the largest ℓ_2-norm and set the remaining rows to zero. Since the estimand θ also has k nonzero rows, applying the triangle inequality yields $\mathbb{E}\|\tilde{\theta} -$

same rate applies to the discretized model. Nevertheless, it is straightforward to extend the arguments in [34], Section 4.1, to show that the rate of $\Psi_{\|\cdot\|}(p,k)$ applies to the discretized model in Section 3 as long as $t = \Omega(\sqrt{\log p})$, independent of the unitarily invariant norm. In particular, the lower bound in [34], Section 4.1.1, applies verbatim due to the data processing inequality of the KL divergence, which is attained by the same estimator defined in [34], Section 4.1.2, if $2^{-t} \le 1/\sqrt{p}$.

$\theta\|_{\mathrm{S}_2}^2 = O(k^2 \log \frac{\mathrm{e}p}{k})$, which implies $\mathbb{E}\|\tilde{\theta} - \theta\|_{\mathrm{S}_q}^2 = O(k^{(2/q+1)\vee 2} \log \frac{\mathrm{e}p}{k})$, since $\|\cdot\|_{\mathrm{S}_q} \le (1 \vee k^{1/q-1/2})\|\cdot\|_{\mathrm{S}_2}$ for all rank-k matrices. Finally, simply estimating θ by the observation X achieves $O(p^{2/q+1})$.

Comparing the minimax rate (43) with the computationally lower and upper bounds (44)–(45), we obtain the following result, assuming Hypothesis 1:

- For $q \in [1, 2]$, using the entrywise thresholding estimator defined above, the minimax rate is attained within a logarithmic factor simultaneously for all k;
- For $q \in (2, \infty]$, the minimax rate $\Psi_{\mathrm{S}_q}(p, k)$ cannot be attained by computationally efficient estimator if $k = \Theta(p^\alpha)$ for all $\alpha \in (0, \frac{q}{2+q})$. In this regime, entrywise thresholding is optimal within a sub-polynomial factor among all randomized polynomial-time procedures.

More generally, one can show that for all *quadratic norms* (see [12], page 95), entrywise thresholding is near optimal (within a sub-polynomial factor) among all computationally efficient estimators. This extends the above result since Schatten-q norm is quadratic if and only if $q \in [2, \infty]$.

6. Proofs.

We present below the proofs of Lemmas 1–4. The proofs of Theorem 1 and Lemmas 5 and 6 are deferred to the supplement [35].

6.1. *Proof of Lemma* 1. Under $\mathbb{P}_0$, $T_{\mathrm{lin}} \sim N(0, 1)$, hence $\mathbb{P}_0\{T_{\mathrm{lin}} > \tau\} = \overline{\Phi}(\tau)$. Under $\mathbb{P}_\theta$ for any $\theta \in \mathcal{M}$, $T_{\mathrm{lin}} \sim N(\bar{\theta}, 1)$, where $\bar{\theta} \triangleq \frac{1}{p}\sum \theta_{ij} \ge \frac{k^2\lambda}{p}$ by the definition of $\mathcal{M}$. Therefore $\mathbb{P}_\theta\{T_{\mathrm{lin}} \le \tau\} \le \overline{\Phi}(\tau)$. Then (10) follows in view of the Chernoff bound $\overline{\Phi}(\tau) \le \frac{1}{2}\exp(-\tau^2/2)$.

By the union bound, $\mathbb{P}_0\{T_{\mathrm{scan}} > \tau'\} \le \binom{p}{k}^2 \mathbb{P}_0\{\sum_{i,j=1}^k X_{ij} > k\tau'\} \le \binom{p}{k}^2 \exp(-\tau'^2/2) \le \exp(-\frac{c}{2}\log\binom{p}{k})$. For any $\theta \in \mathcal{M}$, denote by $U \times V$ the support of θ. Then $|U|, |V| \ge k$. Let I, J be independently and uniformly drawn at random from all subsets of cardinality k of U and V, respectively. Then $\mathbb{E}[\sum_{i\in I, j\in J}\theta_{ij}] \ge \sum_{i\in U, j\in V}\theta_{ij}\mathbb{E}[\mathbf{1}_{\{i\in I\}}\mathbf{1}_{\{j\in J\}}] = \frac{k^2}{|U||V|}\sum_{i\in U, j\in V}\theta_{ij} \ge \lambda k^2$. Therefore there exist $S \subset U$ and $T \subset V$, such that $|S| = |T| = k$ and $\sum_{i\in S, j\in T}\theta_{ij} \ge \lambda k^2$. Then $\sum_{i\in S, j\in T} X_{ij} \sim N(\mu, k^2)$, where $\mu \ge \lambda k^2$. Therefore $\mathbb{P}_\theta\{T_{\mathrm{scan}} \le \tau'\} \le \mathbb{P}_\theta\{\sum_{i\in S, j\in T} X_{ij} \le k\tau'\} \le \exp(-\frac{(\mu - k\tau')_+^2}{2k}) \le \exp(-\frac{1}{2}(\lambda k - \tau')_+^2)$.

The desired bound (12) on $T_{\max}$ follows from analogous arguments since $T_{\max}$ coincides with T_{scan} with parameter $k = 1$.

6.2. *Proof of Lemma* 2. For the first claim, the marginal density function of W is f_1 in (19). So by definition,

$$\begin{aligned}\mathsf{TV}(\mathcal{L}(W), N(\mu, 1)) &= \frac{1}{2}\int_{\mathbb{R}} |f_1(x) - \varphi(x - \mu)|\,\mathrm{d}x = \overline{\Phi}(M - \mu) + \overline{\Phi}(M + \mu)\\ &\le 2\overline{\Phi}(M - \mu) \le \exp(-(M - \mu)^2/2) \le \exp(-(M^2 - 1)/2),\end{aligned}$$

were the last inequality is due to the fact that for any $0 < \mu \le \frac{1}{2M}$, $(M - \mu)^2 \ge (M - \frac{1}{2M})^2 \ge M^2 - 1$. For the second claim, the marginal density function of W is $f = \frac{1}{2}(f_0 + f_1) = c_0 \varphi(x) \mathbf{1}_{\{|x| \le M\}}$. Thus

$$\begin{aligned} \mathsf{TV}(\mathcal{L}(W), N(0,1)) &= \frac{1}{2} \int_{\mathbb{R}} \left| c_0 \varphi(x) \mathbf{1}_{\{|x| \le M\}} - \varphi(x) \right| \mathrm{d}x = 2\overline{\Phi}(M) \\ &\le \exp(-M^2/2). \end{aligned}$$

This completes the proof.

6.3. *Proof of Lemma* 3. We need the following result on the total variation between product distributions.

LEMMA 7. $\mathsf{TV}(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i) \le \sum_{i=1}^n \mathsf{TV}(P_i, Q_i)$.

PROOF. Recall the dual representation of the total variation [40],

$$\mathsf{TV}(P, Q) = \min_{P_{AB}} \{ \mathbb{P}\{A \ne B\} : P_A = P, P_B = Q \} \tag{46}$$

with infimum over all couplings of P and Q. Denote by $P_{A_i B_i}$ the optimal coupling of P_i and Q_i so that $\mathbb{P}\{A_i \ne B_i\} = \mathsf{TV}(P_i, Q_i)$. Then $\prod_{i=1}^n P_{A_i B_i}$ is a coupling between the product measures, and the conclusion follows from the union bound. □

PROOF OF LEMMA 3. Let $\widetilde{B} \in \mathbb{R}^{N_2 \times N_2}$ have i.i.d. $N(0,1)$ entries and be independent of A. Let $\widetilde{X} \in \mathbb{R}^{p \times p}$ be obtained by applying operations (21) and (22) to $\widetilde{B}$ instead of B. Then it is straightforward to verify that $\widetilde{X}$ has i.i.d. $N(0,1)$ entries, that is, $\mathcal{L}(\widetilde{X}) = \mathbb{P}_0$. Hence

$$\begin{aligned} \mathsf{TV}(\mathcal{L}(X), \mathcal{L}(\widetilde{X})) &\le \mathsf{TV}(\mathcal{L}(B), \mathcal{L}(\widetilde{B})) \\ &= \mathsf{TV}\left(\prod_{i,j=1}^{N_2} \mathcal{L}(B_{ij}), \prod_{i,j=1}^{N_2} N(0,1) \right) \qquad (47) \\ &\le \sum_{i,j=1}^{N_2} \mathsf{TV}(\mathcal{L}(B_{ij}), N(0,1)) \le N_2^2 \mathrm{e}^{-M^2/2} = \frac{1}{4N}. \end{aligned}$$

Here, the first inequality is due to the data processing inequality for the total variation [17], the second last inequality is due to Lemma 7 and the last inequality is due to Lemma 2. □

6.4. *Proof of Lemma* 4. Recall that N is even with $N_2 = N/2$. When $A \sim \mathcal{G}(N, 1/2, \kappa)$, let $V \subset [N]$ denote the vertex subset of size κ on which the planted clique in A is supported. For any subset $S \subset \{N_2 + 1, \dots, N\}$, we have $S - N_2 \subset [N_2]$. Further define

$$V_1 = \bigl(V \cap \{N_2 + 1, \dots, N\}\bigr) - N_2, \qquad V_2 = V \cap [N_2]. \tag{48}$$

Then $|V_1| + |V_2| = \kappa$, and the A_0 matrix has all ones on $V_1 \times V_2$ and i.i.d. Bernoulli$(1/2)$ entries elsewhere. Define $h : [N_2] \to [p]$ by

$$h(x) = 1 + (x - 1) \bmod p. \tag{49}$$

For $i = 1, 2$, let

$$U_i = h(V_i). \tag{50}$$

By the definition of X, for each $a, b \in [p]$, we can define sets

$$N_{ab} \triangleq \bigl[h^{-1}(a) \times h^{-1}(b)\bigr] \setminus (V_1 \times V_2), \tag{51}$$

$$T_{ab} \triangleq \bigl[h^{-1}(a) \times h^{-1}(b)\bigr] \cap (V_1 \times V_2). \tag{52}$$

1° We first show that the event

$$E = \bigl\{|U_1| \geq k\bigr\} \cap \bigl\{|U_2| \geq k\bigr\} \tag{53}$$

occurs with high probability. To this end, first note that

$$\begin{aligned} \mathbb{P}\bigl\{|V_1| < \kappa/4\bigr\} &\leq \sum_{j=1}^{\kappa/4} \frac{\binom{N_2}{j}\binom{N_2}{\kappa - j}}{\binom{N}{\kappa}} \leq \frac{\kappa}{4} \frac{\binom{N_2}{\kappa/4}\binom{N_2}{3\kappa/4}}{\binom{N}{\kappa}} = \frac{\kappa}{4} \frac{\binom{\kappa}{\kappa/4}\binom{N-\kappa}{N_2 - \kappa/4}}{\binom{N}{N_2}} \\ &\leq \frac{\kappa}{4}\left(\frac{\mathrm{e}}{4}\right)^{\kappa/4} \sqrt{\frac{2N}{N - \kappa}} \leq \frac{\kappa}{2\sqrt{2}}\left(\frac{\mathrm{e}}{4}\right)^{\kappa/4}, \end{aligned} \tag{54}$$

where the second inequality is due to the fact that $j \mapsto \binom{N_2}{j}\binom{N_2}{\kappa - j}$ is increasing for $j \leq (\kappa - 1)/2$, the third inequality is by the bound on the central binomial coefficient $\frac{2^{2n}}{\sqrt{4n}} \leq \binom{2n}{n} \leq \frac{2^{2n}}{\sqrt{2n}}$ [29], equation (2.12), and $\binom{n}{k} \leq (\frac{\mathrm{e}n}{k})^k$ and the last inequality is due to $N \geq 2\kappa$. By symmetry, since $|V_1| \overset{(\mathrm{d})}{=} |V_2|$ and $|V_1| + |V_2| = \kappa$, $\mathbb{P}\{|V_2| < \kappa/4\} = \mathbb{P}\{|V_1| > 3\kappa/4\}$ also satisfies the upper bound (54).

Note that conditioning on the size $|V_1| = \kappa_1$, the set V_1 is chosen uniformly at random among all κ_1 subsets of $[N_2]$. Thus, for any $\kappa_1 \in [\kappa/4, 3\kappa/4]$ and $c_0 = 1/20$,

$$\begin{aligned} \mathbb{P}\bigl\{|U_1| < c_0\kappa \big| |V_1| = \kappa_1\bigr\} &\leq \sum_{j=\lceil \kappa_1/\ell \rceil}^{c_0\kappa} \frac{\binom{p}{j}\binom{j\ell}{\kappa_1}}{\binom{N_2}{\kappa_1}} \leq c_0\kappa \frac{\binom{p}{c_0\kappa}\binom{c_0\kappa\ell}{\kappa_1}}{\binom{N_2}{\kappa_1}} \\ &\leq c_0\kappa \left(\frac{\mathrm{e}p}{c_0\kappa}\right)^{c_0\kappa} \left(\frac{\mathrm{e}c_0\kappa\ell}{\kappa_1}\right)^{\kappa_1} \left(\frac{\kappa_1}{N_2 - \kappa_1}\right)^{\kappa_1} \end{aligned}$$

$$\le c_0\kappa \exp\biggl(c_0\kappa \log \frac{\mathrm{e}p}{c_0\kappa} - \frac{\kappa}{4}\log\frac{N_2-\kappa}{\mathrm{e}c_0\kappa\ell}\biggr)$$
$$\le \frac{\kappa}{20}\exp\biggl(-\frac{\kappa}{5}\log\frac{p}{\kappa}\biggr).$$

Here the first inequality is because $j \mapsto \binom{p}{j}\binom{j\ell}{\kappa_1}$ is increasing for $j \le (p-1)/2$, the last inequality holds under the assumption that $\kappa \ge 20$ and $p \ge 2\kappa$. Since $k = \lfloor \kappa/20 \rfloor$, the last two displays together lead to

$$\begin{aligned}\mathbb{P}\{|U_1| < k\} &\le \sum_{\kappa_1=0}^{\kappa} \mathbb{P}\{|U_1| < \kappa/20 \,|\, |V_1| = \kappa_1\}\mathbb{P}\{|V_1| = \kappa_1\}\\ &\le \mathbb{P}\{|V_1| < \kappa/4\} + \mathbb{P}\{|V_1| > 3\kappa/4\}\\ &\quad + \max_{\kappa_1\in[\kappa/4,3\kappa/4]} \mathbb{P}\{|U_1| < \kappa/20 \,|\, |V_1| = \kappa_1\}\\ &\le \kappa\biggl(\frac{\mathrm{e}}{4}\biggr)^{\kappa/4} + \frac{\kappa}{20}\exp\biggl(-\frac{\kappa}{5}\log\frac{p}{\kappa}\biggr),\end{aligned}$$

and the union bound further leads to

$$\begin{aligned}\mathbb{P}\{E^c\} &\le 2\mathbb{P}\{|U_1| < k\} \le 2\kappa\biggl(\frac{\mathrm{e}}{4}\biggr)^{\kappa/4} + \frac{\kappa}{10}\exp\biggl(-\frac{\kappa}{5}\log\frac{p}{\kappa}\biggr)\\ &\le 40k\biggl(\frac{\mathrm{e}}{4}\biggr)^{5k} + 2k\exp\biggl(-4k\log\frac{p}{20k}\biggr).\end{aligned} \tag{55}$$

2° Conditioned on V, we generate a random matrix $\widetilde{B} = (\widetilde{B}_{ij}) \in \mathbb{R}^{N_2\times N_2}$ with independent entries where

$$\widetilde{B}_{ij} \sim N(\mu, 1) \quad \text{if } (i,j) \in V_1 \times V_2, \qquad \widetilde{B}_{ij} \sim N(0,1) \quad \text{otherwise.} \tag{56}$$

Then we apply (21) and (22) to $\widetilde{B}$ instead of B to obtain a $p \times p$ random matrix $\widetilde{X}$. The intuition is that $\widetilde{B}$ and $\widetilde{X}$ correspond to the ideal input and output of the reduction scheme, in the sense that the $\mathcal{L}(\widetilde{X})$ is, as we show next, close to a desired mixture on the alternative hypotheses. Our choice of the distribution $\mathcal{F}_0$ and $\mathcal{F}_1$ ensures that B is close to the ideal case $\widetilde{B}$ in total variation, and the data processing inequality guarantees that the output X is also close to $\widetilde{X}$.

To this end, note that conditioned on V, for each $a, b \in [p]$, we have

$$\widetilde{X}_{ab} = \frac{1}{\ell}\sum_{i\in h^{-1}(a), j\in h^{-1}(b)} \widetilde{B}_{ij} = \frac{1}{\ell}\biggl(\sum_{(i,j)\in N_{ab}} \widetilde{B}_{ij} + \sum_{(i,j)\in T_{ab}} \widetilde{B}_{ij}\biggr), \tag{57}$$

where the sets N_{ab} and T_{ab} are defined in (51) and (52), respectively. The last two displays together imply that $\widetilde{X}_{ab}$ follows the Gaussian distribution with unit variance and mean $\mathbb{E}[\widetilde{X}_{ab}] = \frac{\mu|T_{ab}|}{\ell}$. Since for any $(a,b) \in (U_1, U_2)$, $|T_{ab}| \ge 1$, we have $\mathbb{E}[\widetilde{X}_{ab}] \ge \frac{\mu}{\ell} = \frac{2\mu p}{N}$. Last but not the least, since the entries of $\widetilde{X}$ are sums of mutually independent random variables, they are mutually independent themselves. Note that for each fixed V [and hence fixed (V_1, V_2) and (U_1, U_2)], $\mathbf{1}_E$ is deterministic. Therefore, for any V such that $\mathbf{1}_E = 1$, there exists some $\theta = \theta(V) \in \mathcal{M}(p, k, \frac{2\mu p}{N})$ such that $\mathcal{L}(\widetilde{X}|V) = \mathbb{P}_\theta$. Define the probability distribution $\pi = \mathcal{L}(\theta(V)|E)$, which is supported on the set $\mathcal{M}(p, k, \frac{2\mu p}{N})$. Then $\mathcal{L}(\widetilde{X}|E) = \mathbb{P}_\pi$, which is a mixture of distributions of $\{\mathbb{P}_\theta : \theta \in \mathcal{M}(p, k, \frac{2\mu p}{N})\}$ with respect to the prior π.

It remains to show that the law of X is close to the mixture $\mathbb{P}_\pi$. By the convexity of the $(P, Q) \mapsto \mathsf{TV}(P, Q)$, we have

$$
\begin{aligned}
\mathsf{TV}\big(\mathcal{L}(X), \mathcal{L}(\widetilde{X})\big) &\le \mathbb{E}_V\big[\mathsf{TV}\big(\mathcal{L}(X|V), \mathcal{L}(\widetilde{X}|V)\big)\big] \\
&\le \mathbb{E}_V\big[\mathsf{TV}\big(\mathcal{L}(B|V), \mathcal{L}(\widetilde{B}|V)\big)\big] \\
&\le \sum_{i,j=1}^{N_2} \mathsf{TV}\big(\mathcal{L}(B_{ij}|V), \mathcal{L}(\widetilde{B}_{ij}|V)\big) \\
&\le N_2^2 \mathrm{e}^{(M^2-1)/2} \le \frac{\mathrm{e}^{1/2}}{4N} \le \frac{1}{p},
\end{aligned}
\tag{58}
$$

where the second inequality is by the data processing inequality, the third inequality is due to Lemma 7 since conditioned on V both B and $\widetilde{B}$ have independent entries, and the fourth inequality is by Lemma 2, and the last inequality follows from the assumption that $M \ge \sqrt{6 \log N}$. Finally, using $\mathsf{TV}(\mathcal{L}(\widetilde{X}), \mathcal{L}(\widetilde{X}|E)) = \mathbb{P}\{E^{\mathrm{c}}\}$, we obtain

$$
\begin{aligned}
\mathsf{TV}\big(\mathcal{L}(X), \mathbb{P}_\pi\big) &\le \mathsf{TV}\big(\mathcal{L}(X), \mathcal{L}(\widetilde{X})\big) + \mathsf{TV}\big(\mathcal{L}(\widetilde{X}), \mathcal{L}(\widetilde{X}|E)\big) \\
&\le \frac{1}{p} + \mathbb{P}\{E^c\},
\end{aligned}
\tag{59}
$$

where the last inequality is due to (58). In view of (55), this completes the proof.

SUPPLEMENTARY MATERIAL

Supplement to "Computational barriers in minimax submatrix detection" (DOI: 10.1214/14-AOS1300SUPP; .pdf). We provide proofs of Theorem 1 and Lemmas 5 and 6.

REFERENCES

[1] ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

[2] ALON, N., ANDONI, A., KAUFMAN, T., MATULEF, K., RUBINFELD, R. and XIE, N. (2007). Testing k-wise and almost k-wise independence. In *STOC'07—Proceedings of the* 39*th Annual ACM Symposium on Theory of Computing* 496–505. ACM, New York. MR2402475

[3] ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Ninth Annual ACM–SIAM Symposium on Discrete Algorithms* (*San Francisco*, *CA*, 1998) 594–598. ACM, New York. MR1642973

[4] AMES, B. P. W. and VAVASIS, S. A. (2011). Nuclear norm minimization for the planted clique and biclique problems. *Math*. *Program*. **129** 69–89. MR2831403

[5] APPLEBAUM, B., BARAK, B. and WIGDERSON, A. (2010). Public-key cryptography from different assumptions. In *STOC'*10—*Proceedings of the* 2010 *ACM International Symposium on Theory of Computing* 171–180. ACM, New York. MR2743266

[6] ARIAS-CASTRO, E. and VERZELEN, N. (2013). Community detection in random networks. Preprint. Available at arXiv:1302.7099.

[7] ARORA, S. and BARAK, B. (2009). *Computational Complexity*: *A Modern Approach*. Cambridge Univ. Press, Cambridge. MR2500087

[8] BALAKRISHNAN, S., KOLAR, M., RINALDO, A., SINGH, A. and WASSERMAN, L. (2011). Statistical and computational tradeoffs in biclustering. In *NIPS* 2011 *Workshop on Computational Trade-Offs in Statistical Learning*.

[9] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. *Journal of Machine Learning Research*: *Workshop and Conference Proceedings* **30** 1–21.

[10] BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann*. *Statist*. **41** 1780–1815. MR3127849

[11] BHAMIDI, S., DEY, P. S. and NOBEL, A. B. (2012). Energy landscape for large average submatrix detection problems in Gaussian random matrices. Preprint. Available at arXiv:1211.2284.

[12] BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[13] BUTUCEA, C. and INGSTER, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* **19** 2652–2688. MR3160567

[14] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found*. *Comput*. *Math*. **9** 717–772. MR2565240

[15] CHANDRASEKARAN, V. and JORDAN, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **110** E1181–E1190. MR3047651

[16] COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ. MR2239987

[17] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci*. *Math*. *Hungar*. **2** 299–318. MR0219345

[18] DEKEL, Y., GUREL-GUREVICH, O. and PERES, Y. (2011). Finding hidden cliques in linear time with high probability. In *ANALCO*11—*Workshop on Analytic Algorithmics and Combinatorics* 67–75. SIAM, Philadelphia, PA. MR2815485

[19] DESHPANDE, Y. and MONTANARI, A. (2013). Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. Preprint. Available at arXiv:1304.7047.

[20] FEIGE, U. and KRAUTHGAMER, R. (2000). Finding and certifying a large hidden clique in a semirandom graph. *Random Structures Algorithms* **16** 195–208. MR1742351

[21] FEIGE, U. and RON, D. (2010). Finding hidden cliques in linear time. In 21*st International Meeting on Probabilistic*, *Combinatorial*, *and Asymptotic Methods in the Analysis of Algorithms* (*AofA'*10) 189–203. Assoc. Discrete Math. Theor. Comput. Sci., Nancy. MR2735341

[22] FELDMAN, V., GRIGORESCU, E., REYZIN, L., VEMPALA, S. S. and XIAO, Y. (2013). Statistical algorithms and a lower bound for detecting planted cliques. In *STOC'*13—*Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 655–664. ACM, New York. MR3210827

[23] HAZAN, E. and KRAUTHGAMER, R. (2011). How hard is it to approximate the best Nash equilibrium? *SIAM J. Comput.* **40** 79–91. MR2765712

[24] JERRUM, M. (1992). Large cliques elude the Metropolis process. *Random Structures Algorithms* **3** 347–359. MR1179827

[25] JUELS, A. and PEINADO, M. (2000). Hiding cliques for cryptographic security. *Des. Codes Cryptogr.* **20** 269–280. MR1779310

[26] KNUTH, D. E. (1969). *The Art of Computer Programming. Vol.* 2: *Seminumerical Algorithms*. Addison-Wesley, Reading, MA. MR0286318

[27] KOIRAN, P. and ZOUZIAS, A. (2014). Hidden cliques and the certification of the restricted isometry property. *IEEE Trans. Inform. Theory* **60** 4999–5006. MR3245368

[28] KOLAR, M., BALAKRISHNAN, S., RINALDO, A. and SINGH, A. (2011). Minimax localization of structural information in large noisy matrices. *Adv. Neural Inf. Process. Syst.* **24** 909–917.

[29] KOSHY, T. (2009). *Catalan Numbers with Applications*. Oxford Univ. Press, Oxford. MR2526440

[30] KRAUTHGAMER, R., NADLER, B. and VILENCHIK, D. (2013). Do semidefinite relaxations really solve sparse PCA? Preprint. Available at arXiv:1306.3690.

[31] KUVCERA, L. (1992). A generalized encryption scheme based on random graphs. In *Graph-Theoretic Concepts in Computer Science* (*Fischbachau*, 1991). *Lecture Notes in Computer Science* **570** 180–186. Springer, Berlin. MR1245056

[32] KUVCERA, L. (1995). Expected complexity of graph partitioning problems. *Discrete Appl. Math.* **57** 193–212. MR1327775

[33] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. MR0856411

[34] MA, Z. and WU, Y. (2013). Volume ratio, sparsity, and minimaxity under unitarily invariant norms. Preprint. Available at arXiv:1306.3609.

[35] MA, Z. and WU, Y. (2015). Supplement to "Computational barriers in minimax submatrix detection." DOI:10.1214/14-AOS1300SUPP.

[36] RÉNYI, A. (1959). On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hungar.* **10** 193–215 (unbound insert). MR0107575

[37] ROSSMAN, B. (2010). Average-case complexity of detecting cliques. Ph.D. thesis, Massachusetts Institute of Technology. MR2873600

[38] SHABALIN, A. A., WEIGMAN, V. J., PEROU, C. M. and NOBEL, A. B. (2009). Finding large average submatrices in high dimensional data. *Ann. Appl. Stat.* **3** 985–1012. MR2750383

[39] SHIRYAEV, A. N. and SPOKOINY, V. G. (2000). *Statistical Experiments and Decisions: Asymptotic Theory*. *Advanced Series on Statistical Science & Applied Probability* **8**. World Scientific, River Edge, NJ. MR1791434

[40] STRASSEN, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36** 423–439. MR0177430

[41] SUN, X. and NOBEL, A. B. (2013). On the maximal size of large-average and ANOVA-fit submatrices in a Gaussian random matrix. *Bernoulli* **19** 275–294. MR3019495

[42] VERZELEN, N. and ARIAS-CASTRO, E. (2013). Community detection in sparse random networks. Preprint. Available at arXiv:1308.2955.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: zongming@wharton.upenn.edu

DEPARTMENT OF ELECTRICAL
AND COMPUTER ENGINEERING
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
URBANA, ILLINOIS 61801
USA
E-MAIL: yihongwu@illinois.edu