STATISTICAL AND COMPUTATIONAL LIMITS OF DETECTING AND RECOVERING HIDDEN SUBMATRICES

Marom Dadon, Wasim Huleihel, and Tamir Bendory

Tel Aviv University
Department of Electrical Engineering-Systems
Tel Aviv 6997801, Israel

ABSTRACT

We study the problems of detection and recovery of hidden submatrices with elevated means inside a large Gaussian random matrix. We consider two different structures for the planted submatrices. In the first model, the planted matrices are disjoint, and their row and column indices can be arbitrary. Inspired by scientific applications, the second model restricts the row and column indices to be consecutive. In the detection problem, under the null hypothesis, the observed matrix is a realization of independent and identically distributed standard normal entries. Under the alternative, there exists a set of hidden submatrices with elevated means inside the same standard normal matrix. Recovery refers to the task of locating the hidden submatrices. For both problems, and for both models, we characterize the statistical and computational barriers by deriving information-theoretic lower bounds, designing and analyzing algorithms matching those bounds, and proving computational lower bounds based on the low-degree polynomials conjecture.

***Index Terms*—** Signal detection and recovery over networks, computational limits, hidden structures.

1. INTRODUCTION

This paper studies the detection and recovery problems of hidden submatrices inside a large Gaussian random matrix. We consider two statistical models for the planted submatrices. In the first model, the planted matrices are disjoint, and their row and column indices can be arbitrary. The detection and recovery variants of this model are well-known as the *submatrix detection* and *submatrix recovery* (or localization) problems, respectively, and received significant attention in the last few years, e.g., [1–13], and references therein. Specifically, for the case of a *single* planted submatrix, the task is to detect the presence of a small $k \times k$ submatrix with entries sampled from a distribution $\mathcal{P}$ in an $n \times n$ matrix of samples from a distribution $\mathcal{Q}$. In the special case where $\mathcal{P}$ and $\mathcal{Q}$ are Gaussians, the statistical and computational barriers, i.e., information-theoretic lower bounds, algorithms, and computational lower bounds, were studied in great detail and were characterized in [1–4, 7, 13]. When $\mathcal{P}$ and $\mathcal{Q}$ are Bernoulli random variables, the detection task is well-known as the planted dense subgraph problem, which has also been studied extensively in the literature, e.g., [4–6, 12]. Most notably, for both the Gaussian and Bernoulli problems, it is well understood by now that there appears to be a statistical-computational gap between the minimum value of k at which detection can be solved, and the minimum value of k at which detection can be solved in polynomial time (i.e., with an efficient algorithm). The statistical and computational barriers of the recovery problem have also received significant attention in the literature, e.g., [9, 10, 12, 14, 15], covering several types of distributions, as well as single and (non-overlapping) multiple planted submatrices.

The general submatrix model, where the planted column and row indices are arbitrary, might be less realistic in certain scientific and engineering applications. Accordingly, we also analyze a second model that restricts the row and column indices to be consecutive. One important motivation for this model stems from single-particle cryo-electron microscopy (cryo-EM): a leading technology to elucidate the three-dimensional atomic structure of macromolecules, such as proteins [16, 17]. At the beginning of the algorithmic pipeline of cryo-EM, it is required to locate multiple particle images (tomographic projections of randomly oriented copies of the sought molecular structure) in a highly noisy, large image [18, 19]. This task is dubbed particle picking. While many particle picking algorithms were designed, e.g., [20, 21], this work can be seen as a first attempt to unveil the statistical and computational properties of this task that were not analyzed heretofore. **Due to page length limitation we relegate our proofs and some of our discussions to an auxiliary file [22].**

2. PROBLEM FORMULATION

2.1. The detection problem

Let (m, k, n) be three natural numbers, satisfying $m \cdot k \leq n$. We emphasize that the values of m, k, and λ, are allowed to be functions of n—the dimension of the observation. Let $\mathcal{K}_{k,m,n}$ denote all possible sets that can be represented as a

W.H. was supported by the IISF grant 1734/21. T.B. was partially supported by the NSF-BSF grant 2019752, the BSF grant no.2020159, and the ISF grant no.1924/21.

union of m disjoint subsets of $[n]$, each of size k; Formally,

$$\mathcal{K}_{k,m,n} \triangleq \Big\{ \mathsf{K}_{k,m} = \bigcup_{i=1}^{m} \mathsf{S}_i \times \mathsf{T}_i : \ \mathsf{S}_i, \mathsf{T}_i \subset \mathcal{C}_k, \ \forall i \in [m], \\ (\mathsf{S}_i \times \mathsf{T}_i) \cap (\mathsf{S}_j \times \mathsf{T}_j) = \emptyset, \ \forall i \neq j \in [m] \Big\}, \quad (1)$$

where $\mathcal{C}_k \triangleq \{\mathsf{S} \subset [n] : \ |\mathsf{S}| = k\}$, i.e., it is the set of all subsets of $[n]$ of size k. Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian probability measure with mean μ and variance σ^2. We let $\mathcal{Q} = \mathcal{N}(0,1)$ and $\mathcal{P} = \mathcal{N}(\lambda, 1)$, where $\lambda > 0$ is interpreted as the signal-to-noise ratio (SNR) of the underlying model. We shall refer to an element of $\mathcal{K}_{k,m,n}$ as a set of *planted submatrices.*

Definition 1 *Let* $\mathsf{SD}(n,k,m,\mathcal{P},\mathcal{Q})$ *denote the hypothesis testing problem with observation* $\mathsf{X} \in \mathbb{R}^{n \times n}$ *and hypotheses*

$$\mathcal{H}_0 : \mathsf{X} \sim \mathcal{Q}^{\otimes n \times n} \quad \text{vs.} \quad \mathcal{H}_1 : \mathsf{X} \sim \mathcal{D}(n,k,m,\mathcal{P},\mathcal{Q}), \quad (2)$$

where $\mathcal{D}(n,k,m,\mathcal{P},\mathcal{Q})$ *is the distribution of matrices* X *with entries* $\mathsf{X}_{ij} \sim \mathcal{P}$ *if* $i,j \in \mathsf{K}_{k,m}$ *and* $\mathsf{X}_{ij} \sim \mathcal{Q}$ *otherwise that are conditionally independent given* $\mathsf{K}_{k,m}$*, which is chosen uniformly at random over all subsets of* $\mathcal{K}_{k,m,n}$*.*

In some applications, however, the planted submatrices are defined by a set of consecutive rows and a set of consecutive columns (e.g., when those submatrices model images like in cryo-EM). Accordingly, let $\mathcal{K}^{\mathsf{con}}_{k,m,n}$ be defined as $\mathcal{K}_{k,m,n}$ but with $\mathcal{C}_k$ replaced by $\mathcal{C}^{\mathsf{con}}_k \triangleq \{\mathsf{S} \subset [n] : |\mathsf{S}| = k, \mathsf{S} \text{ is consecutive}\}$, i.e., the set of all subsets of $[n]$ of size k with consecutive elements.

Definition 2 *Let* $\mathsf{CSD}(n,k,m,\mathcal{P},\mathcal{Q})$ *denote the hypothesis testing problem with observation* $\mathsf{X} \in \mathbb{R}^{n \times n}$ *and hypotheses*

$$\mathcal{H}_0 : \mathsf{X} \sim \mathcal{Q}^{\otimes n \times n} \quad \text{vs.} \quad \mathcal{H}_1 : \mathsf{X} \sim \widetilde{\mathcal{D}}(n,k,m,\mathcal{P},\mathcal{Q}), \quad (3)$$

where $\widetilde{\mathcal{D}}(n,k,m,\mathcal{P},\mathcal{Q})$ *is the distribution of matrices* X *with entries* $\mathsf{X}_{ij} \sim \mathcal{P}$ *if* $i,j \in \mathsf{K}_{k,m}$ *and* $\mathsf{X}_{ij} \sim \mathcal{Q}$ *otherwise that are conditionally independent given* $\mathsf{K}_{k,m}$*, which is chosen uniformly at random over all subsets of* $\mathcal{K}^{\mathsf{con}}_{k,m,n}$*.*

The difference between $\mathcal{K}_{k,m,n}$ and $\mathcal{K}^{\mathsf{con}}_{k,m,n}$ is depicted in Figure 1; it is evident that the submatrices in $\mathcal{K}_{k,m,n}$ can appear everywhere, while those in $\mathcal{K}^{\mathsf{con}}_{k,m,n}$ are consecutive.

Observing X, a detection algorithm $\mathcal{A}_n$ for the problems above is tasked with outputting a decision in $\{0,1\}$. We define the *risk* of a detection algorithm $\mathcal{A}_n$ as the sum of its Type-I and Type-II errors probabilities, namely,

$$\mathsf{R}(\mathcal{A}_n) \triangleq \mathbb{P}_{\mathcal{H}_0}(\mathcal{A}_n(\mathsf{X}) = 1) + \mathbb{P}_{\mathcal{H}_1}(\mathcal{A}_n(\mathsf{X}) = 0), \quad (4)$$

where $\mathbb{P}_{\mathcal{H}_0}$ and $\mathbb{P}_{\mathcal{H}_1}$ denote the probability distributions under the null hypothesis and the alternative hypothesis, respectively. If $\mathsf{R}(\mathcal{A}_n) \to 0$ as $n \to \infty$, then we say that $\mathcal{A}_n$ solves the detection problem.

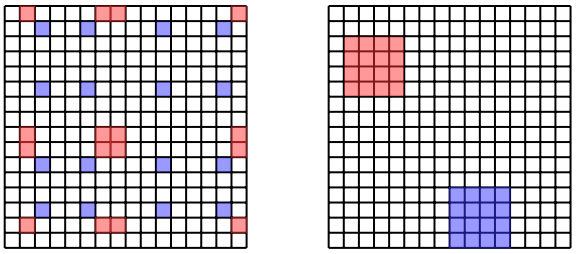

Fig. 1: Illustration of the models considered in this paper: $\mathcal{K}_{k,m,n}$ of Definition 1 (left) and $\mathcal{K}^{\mathsf{con}}_{k,m,n}$ of Definition 2 (right), for $k = 4$, $m = 2$, and $n = 16$.

2.2. The recovery problem

We consider also the recovery variant of the problem in Definition 2. Note that the submatrix recovery problem that corresponds to the model in Definition 1, was investigated in [9]. In the recovery problem, we assume that the data follow the distribution under $\mathcal{H}_1$ in Definition 2, and the inference task is to recover the location of the planted submatrices. This is the analog of the particle picking problem in cryo-EM that was introduced in Section 1.

Definition 3 *Assume that* $\mathsf{X} \in \mathbb{R}^{n \times n} \sim \widetilde{\mathcal{D}}(n,k,m,\mathcal{P},\mathcal{Q})$*, where* $\widetilde{\mathcal{D}}(n,k,m,\mathcal{P},\mathcal{Q})$ *is the distribution of matrices* X *with entries* $\mathsf{X}_{ij} \sim \mathcal{P}$ *if* $i,j \in \mathsf{K}^\star$ *and* $\mathsf{X}_{ij} \sim \mathcal{Q}$ *otherwise that are conditionally independent given* $\mathsf{K}^\star \in \mathcal{K}^{\mathsf{con}}_{k,m,n}$*. The goal is to recover the hidden submatrices* $\mathsf{K}^\star$*, up to a permutation of the submatrices indices, given the matrix* X*. We let* $\mathsf{CSR}(n,k,m,\mathcal{P},\mathcal{Q})$ *denote this recovery problem. We say that* $\hat{\mathsf{K}}$ *achieves exact recovery of* $\mathsf{K}^\star$*, if, asymptotically as* $n \to \infty$*,* $\sup_{\mathsf{K}^\star \in \mathcal{K}^{\mathsf{con}}_{k,m,n}} \mathbb{P}(\hat{\mathsf{K}} \neq \mathsf{K}^\star) \to 0$*.*

Similarly to the detection problem, we will consider both unconstrained and polynomial time algorithms, and we aim to derive necessary and sufficient conditions for when it is impossible and possible to recover the underlying submatrices.

3. MAIN RESULTS

3.1. The detection problem

Upper bounds. Let us propose three algorithms and analyze their performance. Define the statistics, $\mathsf{T}_{\mathsf{sum}}(\mathsf{X}) \triangleq \sum_{i,j \in [n]} \mathsf{X}_{ij}$, $\mathsf{T}^{\mathsf{SD}}_{\mathsf{scan}}(\mathsf{X}) \triangleq \max_{\mathsf{K} \in \mathcal{K}_{k,1,n}} \sum_{i,j \in \mathsf{K}} \mathsf{X}_{ij}$, and $\mathsf{T}^{\mathsf{CSD}}_{\mathsf{scan}}(\mathsf{X}) \triangleq \max_{\mathsf{K} \in \mathcal{K}^{\mathsf{con}}_{k,1,n}} \sum_{i,j \in \mathsf{K}} \mathsf{X}_{ij}$. The sum statistics amounts to adding up all the elements of X, while the scan statistics enumerate all $k \times k$ submatrices of X in $\mathcal{K}_{k,1,n}$ and $\mathcal{K}^{\mathsf{con}}_{k,1,n}$, and take the submatrix with the maximal sum of entries, respectively. Fix $\delta > 0$. Then, our tests are defined as, $\mathcal{A}_{\mathsf{sum}}(\mathsf{X}) \triangleq \mathbb{1}\{\mathsf{T}_{\mathsf{sum}}(\mathsf{X}) \geq \tau_{\mathsf{sum}}\}$, $\mathcal{A}^{\mathsf{SD}}_{\mathsf{scan}}(\mathsf{X}) \triangleq \mathbb{1}\{\mathsf{T}^{\mathsf{SD}}_{\mathsf{scan}}(\mathsf{X}) \geq \tau^{\mathsf{SD}}_{\mathsf{scan}}\}$, and $\mathcal{A}^{\mathsf{CSD}}_{\mathsf{scan}}(\mathsf{X}) \triangleq \mathbb{1}\{\mathsf{T}^{\mathsf{CSD}}_{\mathsf{scan}}(\mathsf{X}) \geq \tau^{\mathsf{CSD}}_{\mathsf{scan}}\}$, where the thresholds are given by $\tau_{\mathsf{sum}} \triangleq \frac{mk^2\lambda}{2}$, $\tau^{\mathsf{SD}}_{\mathsf{scan}} \triangleq \sqrt{(4+\delta)k^2 \log \binom{n}{k}}$, and $\tau^{\mathsf{CSD}}_{\mathsf{scan}} \triangleq$

$\sqrt{(4+\delta)k^2 \log n}$, and correspond roughly to the average between the expected values of each of the statistics above under the null and alternative hypotheses. The following result provides sufficient conditions under which the risk of each of the above tests is asymptotically small.

Theorem 1 *Consider the detection problems in Definitions 1 and 2. Then, we have the following bounds:*

1. *(Efficient* SD*) There exists an efficient algorithm* $\mathcal{A}_{\mathsf{sum}}$*, such that if* $\lambda = \omega\left(\frac{n}{mk^2}\right)$*, then* $\mathsf{R}\left(\mathcal{A}_{\mathsf{sum}}\right) \to 0$*, as* $n \to \infty$*, for the problems in Definitions 1 and 2.*

2. *(Exhaustive* SD*) There exists an algorithm* $\mathcal{A}^{\mathsf{SD}}_{\mathsf{scan}}$*, such that if* $\lambda = \omega(\sqrt{k^{-1}\log\frac{n}{k}})$*, then* $\mathsf{R}\left(\mathcal{A}^{\mathsf{SD}}_{\mathsf{scan}}\right) \to 0$*, as* $n \to \infty$*, for the problem in Def. 1.*

3. *(Efficient* CSD*) There exists an efficient algorithm* $\mathcal{A}^{\mathsf{CSD}}_{\mathsf{scan}}$*, such that if* $\lambda = \omega(k^{-1}\sqrt{\log\frac{n}{k}})$*, then* $\mathsf{R}(\mathcal{A}^{\mathsf{CSD}}_{\mathsf{scan}}) \to 0$*, as* $n \to \infty$*, for the problem in Def. 2.*

Lower bounds. Recall that the optimal testing error probability is determined by the total variation distance between the distributions under the null and the alternative hypotheses as follows (see, e.g., [23, Lemma 2.1]),

$$\min_{\mathcal{A}_n:\mathbb{R}^{n\times n}\to\{0,1\}} \mathsf{R}(\mathcal{A}_n) = 1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}). \tag{5}$$

Theorem 2 *We have the following results:*

1. *Consider the detection problem in Def. 1. If* $\lambda = o(\frac{n}{mk^2} \wedge \frac{1}{\sqrt{k}})$*, then* $d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) = o(1)$*.*

2. *Consider the detection problem in Def. 2. If* $\lambda = o\left(k^{-1}\right)$*, then* $d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) = o(1)$*.*

Theorem 2 above shows that our upper bounds in Theorem 1 are tight up to poly-log factors. Indeed, item 1 in Theorem 2 complements items 1-2 in Theorem 1, for the SD problem, while item 2 in Theorem 2 complements item 3 in Theorem 1, for the CSD problem. We next give evidence that, based on the low-degree polynomial conjecture, efficient algorithms that run in poly-time do not exist in the regime where the scan test succeeds while the sum test fails.

Computational lower bounds. The premise of the *low-degree polynomials* method is to take low-degree multivariate polynomials in the entries of the observations as a proxy for efficiently-computable functions. Roughly speaking, it takes the projection of likelihood ratio defined by $\mathsf{L}_n \triangleq \mathbb{P}_{\mathcal{H}_1}/\mathbb{P}_{\mathcal{H}_0}$ onto D-degree polynomial space as a proxy to all efficient (poly-logarithmic time) algorithms. Accordingly, detection is possible when $\mathsf{L}_n^{\le \mathsf{D}}$ is unbounded.

Theorem 3 (Computational lower bound) *Consider the detection problem in Definition 1. Then, if* λ *is such that* $\frac{1}{\sqrt{k}} \ll \lambda \ll \frac{n}{mk^2}$*, then* $\left\|\mathsf{L}_n^{\le \mathsf{D}}\right\|_{\mathcal{H}_0} \le O(1)$*, for any* $\mathsf{D} = \Omega(\log n)$*. On the other hand, if* λ *is such that* $\lambda \gg \frac{n}{mk^2}$*, then* $\left\|\mathsf{L}_n^{\le \mathsf{D}}\right\|_{\mathcal{H}_0} \ge \omega(1)$*.*

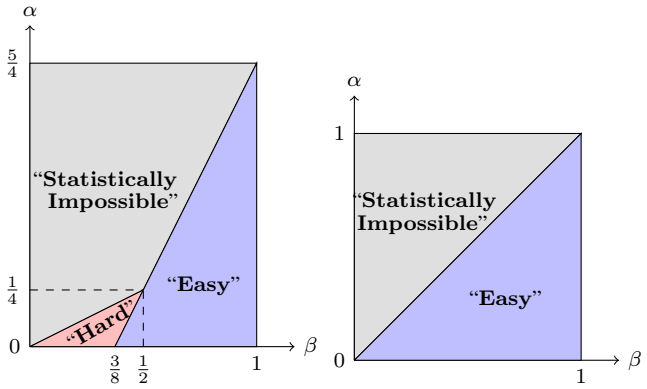

Fig. 2: Phase diagram for SD (left) and CSD (right), as a function of $k = \Theta(n^\beta)$, and $\lambda = \Theta(n^{-\alpha})$, for $m = \Theta(n^{1/4})$ (left) and any $m \ge 1$ (right).

Together with low-degree polynomials conjecture, Theorem 3 implies that if we take degree-$\log n$ polynomials as a proxy for all efficient algorithms, our calculations predict that an $n^{O(\log n)}$ algorithm does not exist when $\frac{1}{\sqrt{k}} \ll \lambda \ll \frac{n}{mk^2}$. This is summarized in the following corollary.

Corollary 4 *Consider the detection problem in Definition 1, and assume that the low-degree polynomials conjecture holds. An* $n^{O(\log n)}$ *algorithm that achieves strong detection does not exist if* λ *is such that* $\frac{1}{\sqrt{k}} \ll \lambda \ll \frac{n}{mk^2}$*.*

These predictions agree precisely with the previously established statistical-computational tradeoffs in the previous subsections. We note that numerical and theoretical evidence for the existence of computational-statistical gaps were observed in other statistical models that are also inspired by cryo-EM, including heterogeneous multi-reference alignment [24, 25] and sparse multi-reference alignment [26].

One way to present the results in Theorems 1–3 is by drawing "easy-hard-impossible" phase transition diagram. Specifically, treating k and λ as polynomials in n, i.e., $k = \Theta(n^\beta)$ and $\lambda = \Theta(n^{-\alpha})$, for some $\alpha \in (0,1)$ and $\beta \in (0,1)$, the statistical and computational feasibility of the SD and CSD problems are demonstrated in Fig. 2, for $m = \Theta(n^{1/4})$ and any $m \ge 1$, respectively. It can be seen that for SD the (α, β)-plane is divided into three regions: statistically impossible, hard, and easy, while for CSD there are only two regions where the problem is either statistically impossible or easy to solve. Interestingly, while it is well-known that the number of planted submatrices m does not play any significant role in the statistical and computational barriers in the submatrix recovery problem, it can be seen that this is not the case for the submatrix detection problem. Finally, to get a sense of how practical results behave alongside with the theory we also present a simulated phase diagram that corresponds to the SD setting in Fig. 2. Specifically, in our simulations we chose $n = 10^9$, $m = n^{1/4}$, and for each one of the 10^2 Monte-Carlo simulations, we generated the $n \times n$ normally distributed random matrices with and without the planted submatrices, over a grid of parameters (k, λ). We then examined the risk of the

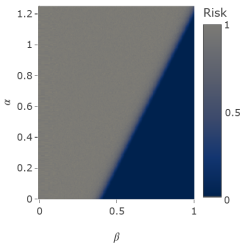

Fig. 3: Simulated phase diagram for SD as a function of $k = \Theta(n^{\beta})$, and $\lambda = \Theta(n^{-\alpha})$, for $m = \Theta(n^{1/4})$.

efficient sum test averaged over the 10^2 Monte-Carlo simulations. It can be seen that the blue region (right triangular) in Fig. 2, where the sum test achieves small risk, coincides with the "easy" region in Fig. 2. A similar figure can be obtained for the CSD setting in Fig. 2.

3.2. The recovery problem

Upper bound. It can be shown that the maximum-likelihood estimator (MLE) , minimizing the error probability, is given by $\hat{\mathsf{K}}_{\mathsf{ML}}(\mathsf{X}) = \arg\max_{\mathsf{K}\in\mathcal{K}^{\mathsf{con}}_{k,m,n}} \sum_{(i,j)\in\mathsf{K}} \mathsf{X}_{ij}$. The computational complexity of MLE is of order n^{2m}. Thus, for $m = O(1)$, the MLE runs in polynomial time, and thus, is efficient. However, if $m = \omega(1)$ then the exhaustive search is not efficient anymore. Nonetheless, the following straightforward modification provably achieves the same asymptotic performance of the MLE above, and at the same time computationally efficient. Before we present this algorithm, we make a simplifying technical assumption on the possible set of planted submatrices. We explain in [22] how this assumption can be removed. *We assume that each pair of submatrices in the underlying planted submatrices* $\mathsf{K}^\star$ *are at least* k *columns and rows far way*. In other words, there are at least k columns and k rows separating any pair of submatrices in $\mathsf{K}^\star$. Similar assumptions are frequently taken when analyzing statistical models inspired by cryo-EM, see, for example [27]. We will refer to the above as the *separation assumption*.

Our recovery algorithm works as follows: in the $\ell \in [m]$ step, we find the ML estimate of a single submatrix using,

$$\hat{\mathsf{K}}_\ell(\mathsf{X}^{(\ell)}) = \underset{\mathsf{K}\in\mathcal{K}^{\mathsf{con}}_{k,1,n}}{\arg\max} \sum_{(i,j)\in\mathsf{K}} \mathsf{X}^{(\ell)}_{ij}, \tag{6}$$

where $\mathsf{X}^{(\ell)}$ is defined recursively as follows: $\mathsf{X}^{(1)} \triangleq \mathsf{X}$, and for $\ell \geq 2$, $\mathsf{X}^{(\ell)} = \mathsf{X}^{(\ell-1)} \odot \mathsf{E}(\hat{\mathsf{K}}_{\ell-1})$, where $\mathsf{E}(\hat{\mathsf{K}}_{\ell-1})$ is an $n \times n$ matrix such that $[\mathsf{E}(\hat{\mathsf{K}}_{\ell-1})]_{ij} = -\infty$, for $(i,j) \in \hat{\mathsf{K}}_{\ell-1}$, and $[\mathsf{E}(\hat{\mathsf{K}}_{\ell-1})]_{ij} = 1$, otherwise. To wit, in each step of the algorithm we "peel" the set of estimated indices (or, estimated submatrices) in previous steps from the search space. This is done by setting the corresponding entries of X to $-\infty$ so that the sum in (6) will not be maximized by previously chosen sets of indices. We denote by $\hat{\mathsf{K}}_{\mathsf{peel}}(\mathsf{X}) = \{\hat{\mathsf{K}}_\ell\}_{\ell=1}^m$ the output of the above peeling algorithm.

Type	Impossible	Hard	Easy
SD	$\lambda \ll \frac{n}{mk^2} \wedge \frac{1}{\sqrt{k}}$	$\frac{n}{mk^2} \wedge \frac{1}{\sqrt{k}} \ll \lambda \ll 1 \wedge \frac{n}{mk^2}$	$\lambda \gg 1 \wedge \frac{n}{mk^2}$
SR	$\lambda \ll \frac{1}{\sqrt{k}}$	$\frac{1}{\sqrt{k}} \ll \lambda \ll 1 \wedge \frac{\sqrt{n}}{k}$	$\lambda \gg 1 \wedge \frac{\sqrt{n}}{k}$
CSD	$\lambda \ll \frac{1}{k}$	NO	$\lambda \gg \frac{1}{k}$
CSR	$\lambda \ll \frac{1}{\sqrt{k}}$	NO	$\lambda \gg \frac{1}{\sqrt{k}}$

Table 1: Statistical and computational thresholds for submatrix detection (SD), submatrix recovery (SR), consecutive submatrix detection (CSD), and consecutive submatrix recovery (CSR), up to poly-log factors. The bounds in the first row for the special case of $m = 1$ and the second row, are known in the literature (e.g., [4, 7, 9, 10]).

Theorem 5 *Consider the recovery problem in Definition 3, and let* C *be a universal constant. If* $\liminf_{n\to\infty} \frac{\lambda}{\sqrt{\mathsf{C}k^{-1}\log n}} > 1$*, then exact recovery is possible via the MLE/peeling algorithm.*

Lower bound. The following result shows that under certain conditions, exact recovery is impossible.

Theorem 6 *Consider the recovery problem in Definition 3. If* $\lambda < \mathsf{C}\sqrt{\frac{\log m}{k}}$*, then exact recovery is impossible, i.e.,* $\inf_{\hat{\mathsf{K}}} \sup_{\mathsf{K}^\star\in\mathcal{K}^{\mathsf{con}}_{k,m,n}} \mathbb{P}[\hat{\mathsf{K}}(\mathsf{X}) \neq \mathsf{K}^\star] > \frac{1}{2}$*, where the infimum ranges over all measurable functions of the matrix* X*.*

Note that there is a gap between detection and exact recovery; the barrier for λ for the former is at k^{-1}, while for the latter at $k^{-1/2}$. In the context of cryo-EM, this indicates a gap between the ability to detect the existence of particle images in the data set, and the ability to perform successful particle picking (exact recovery). Recently, new computational methods were devised to elucidate molecular structures without particle picking, thus bypassing the limit of exact recovery, allowing constructing structures in very low SNR environments, e.g., [27–29]. This in turn opens the door to recovering small molecular structures that induce low SNR [30]. Finally, we summarize our main results in Table 1.

4. CONCLUSIONS AND OUTLOOK

In this paper, we studied the computational and statistical boundaries of the submatrix and consecutive submatrix detection and recovery problems. For both models, we derived asymptotically tight lower and upper bounds on the thresholds for detection and recovery. There are several exciting directions for future work. First, it would be interesting to generalize our results to any pair of distributions $\mathcal{P}$ and $\mathcal{Q}$. In our paper, we assume that the elements inside the planted submatrices are i.i.d., however, it is of practical interest to generalize this assumption and consider the case of dependent entries, e.g., Gaussians with a general covariance matrix. For example, this is the typical statistical model of cryo-EM data [19]. Finally, it will be interesting to prove a computational lower bound for the submatrix recovery problem, e.g., using the recent framework of low-degree polynomials for recovery.

5. REFERENCES

[1] Andrey A Shabalin, Victor J Weigman, Charles M Perou, Andrew B Nobel, et al., "Finding large average submatrices in high dimensional data," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 985–1012, 2009.

[2] Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, and Aarti Singh, "Minimax localization of structural information in large noisy matrices," in *Advances in Neural Information Processing Systems*, 2011, pp. 909–917.

[3] Sivaraman Balakrishnan, Mladen Kolar, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman, "Statistical and computational tradeoffs in biclustering," in *NIPS 2011 workshop on computational trade-offs in statistical learning*, 2011, vol. 4.

[4] Cristina Butucea and Yuri I Ingster, "Detection of a sparse submatrix of a high-dimensional noisy matrix," *Bernoulli*, vol. 19, no. 5B, pp. 2652–2688, 2013.

[5] Ery Arias-Castro and Nicolas Verzelen, "Community detection in dense random networks," *The Annals of Statistics*, vol. 42, no. 3, pp. 940–969, 2014.

[6] Bruce Hajek, Yihong Wu, and Jiaming Xu, "Computational lower bounds for community detection on random graphs," in *Proceedings of The 28th Conference on Learning Theory*, 03–06 Jul 2015, vol. 40, pp. 899–928.

[7] Zongming Ma and Yihong Wu, "Computational barriers in minimax submatrix detection," *Annals of Statistics*, vol. 43, no. 3, pp. 1089–1116, 2015.

[8] Shankar Bhamidi, Partha Dey, and Andrew Nobel, "Energy landscape for large average submatrix detection problems in gaussian random matrices," *Probability Theory and Related Fields*, vol. 168, 08 2017.

[9] Yudong Chen and Jiaming Xu, "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices," *Journal of Machine Learning Research*, vol. 17, no. 27, pp. 1–57, 2016.

[10] Tony Cai, Tengyuan Liang, and Alexander Rakhlin, "Computational and statistical boundaries for submatrix localization in a large noisy matrix," *Annals of Statistics*, vol. 45, no. 4, pp. 1403–1430, 08 2017.

[11] Wasim Huleihel, "Inferring hidden structures in random graphs," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 8, pp. 855–867, 2022.

[12] Matthew Brennan, Guy Bresler, and Wasim Huleihel, "Reducibility and computational lower bounds for problems with planted sparse structure," in *Proceedings of the 31st Conference On Learning Theory*, 06–09 Jul 2018, vol. 75, pp. 48–166.

[13] Matthew Brennan, Guy Bresler, and Wasim Huleihel, "Universality of computational lower bounds for submatrix detection," in *Proceedings of the Thirty-Second Conference on Learning Theory*, 25–28 Jun 2019, vol. 99, pp. 417–468.

[14] Utkan Onur Candogan and Venkat Chandrasekaran, "Finding planted subgraphs with few eigenvalues using the schur–horn relaxation," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 735–759, 2018.

[15] Bruce Hajek, Yihong Wu, and Jiaming Xu, "Information limits for recovering a hidden community," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4729–4745, 2017.

[16] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres, "How cryo-EM is revolutionizing structural biology," *Trends in biochemical sciences*, vol. 40, no. 1, pp. 49–57, 2015.

[17] Dmitry Lyumkis, "Challenges and opportunities in cryo-EM single-particle analysis," *Journal of Biological Chemistry*, vol. 294, no. 13, pp. 5181–5197, 2019.

[18] Amit Singer, "Mathematics for cryo-electron microscopy," in *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*. World Scientific, 2018, pp. 3995–4014.

[19] Tamir Bendory, Alberto Bartesaghi, and Amit Singer, "Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities," *IEEE signal processing magazine*, vol. 37, no. 2, pp. 58–76, 2020.

[20] Feng Wang, Huichao Gong, Gaochao Liu, Meijing Li, Chuangye Yan, Tian Xia, Xueming Li, and Jianyang Zeng, "DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM," *Journal of structural biology*, vol. 195, no. 3, pp. 325–336, 2016.

[21] Ayelet Heimowitz, Joakim Andén, and Amit Singer, "APPLE picker: Automatic particle picking, a low-effort cryo-EM framework," *Journal of structural biology*, vol. 204, no. 2, pp. 215–227, 2018.

[22] Marom Dadon, Wasim Huleihel, and Tamir Bendory, "Detection and recovery of hidden submatrices," *arXiv preprint arXiv:2306.06643*, 2023.

[23] Alexandre B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer Publishing Company, Incorporated, 1st edition, 2008.

[24] Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer, "Heterogeneous multireference alignment: A single pass approach," in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2018, pp. 1–6.

[25] Alexander Spence Wein, *Statistical estimation in the presence of group actions*, Ph.D. thesis, Massachusetts Institute of Technology, 2018.

[26] Tamir Bendory, Oscar Mickelin, and Amit Singer, "Sparse multi-reference alignment: Sample complexity and computational hardness," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8977–8981.

[27] Tamir Bendory, Nicolas Boumal, William Leeb, Eitan Levin, and Amit Singer, "Toward single particle reconstruction without particle picking: Breaking the detection limit," *SIAM Journal on Imaging Sciences*, vol. 16, no. 2, pp. 886–910, 2023.

[28] Shay Kreymer and Tamir Bendory, "Two-dimensional multi-target detection: An autocorrelation analysis approach," *IEEE Transactions on Signal Processing*, vol. 70, pp. 835–849, 2022.

[29] Shay Kreymer, Amit Singer, and Tamir Bendory, "A stochastic approximate expectation-maximization for structure determination directly from cryo-EM micrographs," *arXiv preprint arXiv:2303.02157*, 2023.

[30] Richard Henderson, "The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules," *Quarterly reviews of biophysics*, vol. 28, no. 2, pp. 171–193, 1995.