Proceedings of Machine Learning Research vol 125:1–200, 2020 33rd Annual Conference on Learning Theory

Reducibility and Statistical-Computational Gaps from Secret Leakage

Matthew Brennan BRENNANM@MIT.EDU
Massachusetts Institute of Technology. Department of EECS.

Guy Bresler GUY@MIT.EDU
Massachusetts Institute of Technology. Department of EECS.

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Inference problems with conjectured statistical-computational gaps are ubiquitous throughout modern statistics, computer science, statistical physics and discrete probability. While there has been success evidencing these gaps from the failure of restricted classes of algorithms, progress towards a more traditional reduction-based approach to computational complexity in statistical inference has been limited. These average-case problems are each tied to a different natural distribution, high-dimensional structure and conjecturally hard parameter regime, leaving reductions among them technically challenging. Despite a flurry of recent success in developing such techniques, existing reductions have largely been limited to inference problems with similar structure – primarily mapping among problems representable as a sparse submatrix signal plus a noise matrix, which is similar to the common starting hardness assumption of planted clique (PC).

The insight in this work is that a slight generalization of the planted clique conjecture – secret leakage planted clique (PC_ρ), wherein a small amount of information about the hidden clique is revealed – gives rise to a variety of new average-case reduction techniques, yielding a web of reductions relating statistical problems with very different structure. Based on generalizations of the planted clique conjecture to specific forms of PC_ρ, we deduce tight statistical-computational tradeoffs for a diverse range of problems including robust sparse mean estimation, mixtures of sparse linear regressions, robust sparse linear regression, tensor PCA, variants of dense k-block stochastic block models, negatively correlated sparse PCA, semirandom planted dense subgraph, detection in hidden partition models and a universality principle for learning sparse mixtures. This gives the first reduction-based evidence for a number of conjectured statistical-computational gaps.

We introduce a number of new average-case reduction techniques that also reveal novel connections to combinatorial designs based on the incidence geometry of $\mathbb{F}_r^t$ and to random matrix theory. In particular, we show a convergence result between Wishart and inverse Wishart matrices that may be of independent interest. The specific hardness conjectures for PC_ρ implying our statistical-computational gaps all are in correspondence with natural graph problems such as k-partite, bipartite and hypergraph variants of PC. Hardness in a k-partite hypergraph variant of PC is the strongest of these conjectures and sufficient to establish all of our computational lower bounds. We also give evidence for our PC_ρ hardness conjectures from the failure of low-degree polynomials and statistical query algorithms. Our work raises a number of open problems and suggests that previous technical obstacles to average-case reductions may have arisen because planted clique is not the right starting point. An expanded set of hardness assumptions, such as PC_ρ, may be a key first step towards a more complete theory of reductions among statistical problems.

Keywords: Statistical-computational tradeoffs, average-case complexity, average-case reductions, planted clique, secret leakage

© 2020 M. Brennan & G. Bresler.

1. Introduction

Computational complexity has become a central consideration in statistical inference as focus has shifted to high-dimensional structured problems. In many of these problems, the number of samples or level of signal information theoretically required to solve the problem often is far lower than that required by efficient algorithms. This phenomenon, referred to as a *statistical-computational gap*, was first observed to exist more than two decades ago (Valiant, 1984; Servedio, 1999; Decatur et al., 2000) but only recently has emerged as a trend ubiquitous in problems throughout modern statistics, computer science, statistical physics and discrete probability (Bottou and Bousquet, 2008; Chandrasekaran and Jordan, 2013; Jordan and Mitchell, 2015).

Because statistical inference problems are formulated with probabilistic models on the observed data, there are natural barriers to basing their computational complexity as average-case problems on worst-case complexity assumptions such as P $\neq$ NP (Feigenbaum and Fortnow, 1993; Bogdanov and Trevisan, 2006b; Applebaum et al., 2008). To cope with this complication, a number of different approaches have emerged to provide evidence for conjectured statistical-computational gaps. These can be roughly classified into two categories:

1. **Failure of Classes of Algorithms:** Showing that powerful classes of polynomial-time algorithms fail up to the conjectured computational limit of a problem.

2. **Average-Case Reductions:** The traditional complexity-theoretic approach of exhibiting polynomial time reductions relating statistical-computational gaps in problems to one another.

The line of research providing evidence for statistical-computational gaps through the failure of powerful classes of algorithms has seen a lot of progress in the past few years. There are now established techniques to show lower bounds for average-case problems against a number of classes of algorithms including the sum of squares (SOS) hierarchy (Barak et al., 2016), low-degree polynomials (Hopkins, 2018; Kunisky et al., 2019), statistical query (SQ) algorithms (Feldman et al., 2013; Diakonikolas et al., 2017), classes of circuits (Rossman, 2008), local algorithms (Gamarnik and Sudan, 2017) and message-passing algorithms (Zdeborová and Krzakala, 2016). Further background on this line of work can be found in Section A.1.

While there has been success analyzing barriers to these classes of algorithms, progress towards the reduction-based approach has been more limited. Reductions between average-case problems are more constrained and overall very different from reductions between worst-case problems. As discussed in Barak (2017) and Goldreich (2011), average-case reductions are notoriously delicate and there is a lack of available techniques. Since statistical inference problems are parameterized and only hard in certain parameter regimes, reductions among them are even more constrained. In Section A.2, we discuss general criteria that these reductions need to satisfy to show strong lower bounds. Although technically difficult to obtain, average-case reductions have a number of advantages – by directly transferring hardness between problems, they are future-proof against new classes of algorithms and reveal insights into how the parameters, hidden structures and noise models in these problems correspond to one another.

Despite these challenges, there has been a flurry of recent success in developing techniques for average-case reductions among statistical problems. Since the seminal paper of Berthet and Rigollet (2013a) showing that a statistical-computational gap for a distributionally-robust formulation of sparse PCA follows from the planted clique (PC) conjecture, there have been a number of further reductions. Reductions from PC have been used to show lower bounds for RIP certification (Wang

et al., 2016a; Koiran and Zouzias, 2014), biclustering detection and recovery (Ma and Wu, 2015; Cai et al., 2015a; Cai and Wu, 2018; Brennan et al., 2019a), planted dense subgraph (Hajek et al., 2015; Brennan et al., 2019a), testing k-wise independence (Alon et al., 2007), matrix completion (Chen, 2015) and sparse PCA (Berthet and Rigollet, 2013b,a; Wang et al., 2016b; Gao et al., 2017; Brennan and Bresler, 2019). Several reduction techniques were introduced in (Brennan et al., 2018), providing the first web of average-case reductions among a number of problems involving sparsity. There also have been a number of average-case reductions in the literature starting with different assumptions than the PC conjecture. Hardness conjectures for random CSPs have been used to show hardness in improper learning complexity (Daniely et al., 2014) and hardness of approximation (Feige, 2002). Recent reductions also map from a 3-uniform hypergraph variant of the PC conjecture to SVD for random 3-tensors (Zhang and Xia, 2017) and between learning two-layer neural networks and tensor decomposition (Mondelli and Montanari, 2018a).

A common criticism to the reduction-based approach to computational complexity in statistical inference is that, while existing reductions have introduced nontrivial techniques for mapping precisely between different natural distributions, they are not yet capable of transforming between problems with dissimilar *high-dimensional structures*. In particular, the vast majority of the reductions referenced above map among problems representable as a *sparse submatrix signal plus a noise matrix*, which is similar to the common starting hardness assumption PC. Such a barrier would be fatal to a satisfying reduction-based theory of statistical-computational gaps, as the zoo of statistical problems with gaps contains a broad range of very different high-dimensional structures. This leads directly to the following central question that we aim to address in this work.

Question 1.1 *Can statistical-computational gaps in problems with different high-dimensional structures be related to one another through average-case reductions?*

1.1. Overview

The main objective of this paper is to provide the first evidence that relating differently structured statistical problems through reductions is possible. We show that mild generalizations of the PC conjecture to k-partite and bipartite variants of PC are naturally suited to a number of new average-case reduction techniques. These techniques break out of the sparse submatrix plus noise structure that seemed to constrain prior reductions. They thus show that revealing a tiny amount of information about the hidden clique vertices substantially increases the reach of the reductions approach, providing the first web of reductions among statistical problems with significantly different structure. Our techniques also yield reductions beginning from hypergraph variants of PC which, along with the k-partite and bipartite variants mentioned above, can be unified under a single assumption that we introduce – the secret leakage planted clique (PC_ρ) conjecture. This conjecture makes a precise prediction of what information about the hidden clique can be revealed while PC remains hard.

A summary of our web of average-case reductions is shown in Figure 1. Our reductions yield tight statistical-computational gaps for a range of differently structured problems, including robust sparse mean estimation, variants of dense stochastic block models, detection in hidden partition models, semirandom planted dense subgraph, negatively correlated sparse PCA, mixtures of sparse linear regressions, robust sparse linear regression, tensor PCA and a universality principle for learning sparse mixtures. This gives the first reduction-based evidence supporting a number of gaps observed in the literature (Li, 2017; Balakrishnan et al., 2017a; Diakonikolas et al., 2017; Chen and Xu, 2016; Hajek et al., 2015; Brennan et al., 2018; Fan et al., 2018; Liu et al., 2018; Richard and

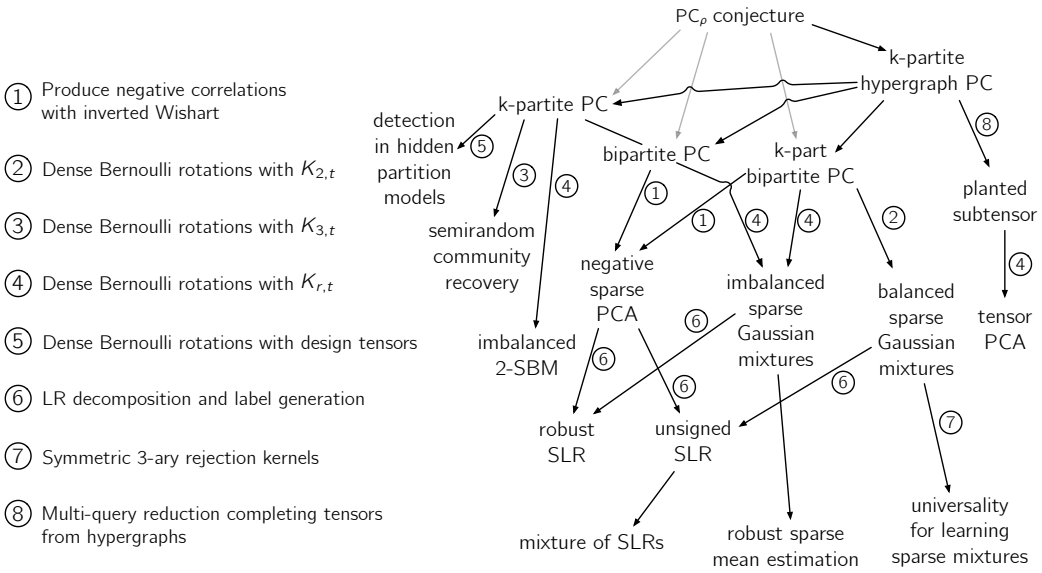

Figure 1: The web of reductions carried out in this paper. An edge indicates existence of a reduction transferring computational hardness from the tail to the head. Edges are labeled with associated techniques and unlabelled edges correspond to simple reductions or specializing a problem to a particular case.

Montanari, 2014; Hopkins et al., 2015; Wein et al., 2019; Azizyan et al., 2013; Verzelen and Arias-Castro, 2017). The specific instantiations of the PC$_\rho$ conjecture needed to obtain these lower bounds correspond to natural k-partite, bipartite and hypergraph variants of PC. Among these hardness assumptions, we show that hardness in a k-partite hypergraph variant of PC (k-HPC^s) is the strongest and sufficient to establish all of our computational lower bounds. We also give evidence for our hardness conjectures from the failure of low-degree polynomials and statistical query algorithms.

Our results suggest that PC may not be the right starting point for average-case reductions among statistical problems. However, surprisingly mild generalizations of PC are all that are needed to break beyond the structural constraints of previous reductions. Generalizing to either PC$_\rho$ or k-HPC^s unifies all of our reductions under a single hardness assumption, now capturing reductions to a range of dissimilarly structured problems including supervised learning tasks and problems over tensors. This suggests PC$_\rho$ and k-HPC^s are both much more powerful candidate starting points than PC. Although we often will focus on providing evidence for statistical-computational gaps, we emphasize that our main contribution is more general – our reductions give a new set of techniques for relating differently structured statistical problems that seem likely to have further applications.

The rest of the paper is structured as follows. In Section 2, we introduce the PC$_\rho$ conjecture and the specific instantiations of this conjecture that imply our computational lower bounds, such as k-HPC^s. In Section 3 we formally introduce the problems in Figure 1 and state our main theorems. In Section 4, we give a very brief overview of our techniques. In Part I, we give further background, more detailed discussions of our main results and technical contributions and state several future directions. In Part II, we formally introduce our main reduction techniques. Part III is devoted to our hardness assumptions and the PC$_\rho$ conjecture, our reductions to the problems in Figure 1 and deducing our main theorems. Reading Part I, Section K and the pseudocode for our reductions gives an accurate summary of the theorems and ideas in this work.

2. Planted Clique and Secret Leakage

In this section, we introduce planted clique and our generalization of the planted clique conjecture. In the planted clique problem (PC), the task is to find the vertex set of a k-clique planted uniformly at random in an n-vertex Erdős-Rényi graph G. Planted clique can equivalently be formulated as a testing problem $\text{PC}(n, k, 1/2)$ (Alon et al., 2007) between the two hypotheses

$$H_0 : G \sim \mathcal{G}(n, 1/2) \quad \text{and} \quad H_1 : G \sim \mathcal{G}(n, k, 1/2)$$

where $\mathcal{G}(n, 1/2)$ denotes the n-vertex Erdős-Rényi graph with edge density $1/2$ and $\mathcal{G}(n, k, 1/2)$ the distribution resulting from planting a k-clique uniformly at random in $\mathcal{G}(n, 1/2)$. This problem can be solved in quasipolynomial time by searching through all vertex subsets of size $(2+\epsilon)\log_2 n$ if $k > (2+\epsilon)\log_2 n$. The *Planted Clique Conjecture* is that there is no polynomial time algorithm solving $\text{PC}(n, k, 1/2)$ if $k = o(\sqrt{n})$.

There is a plethora of evidence in the literature for the PC conjecture. Spectral algorithms, approximate message passing, semidefinite programming, nuclear norm minimization and several other polynomial-time combinatorial approaches all appear to fail to solve PC exactly when $k = o(\sqrt{n})$ (Alon et al., 1998; Feige and Krauthgamer, 2000; McSherry, 2001; Feige and Ron, 2010; Ames and Vavasis, 2011; Dekel et al., 2014; Deshpande and Montanari, 2015a; Chen and Xu, 2016). Lower bounds against low-degree sum of squares relaxations (Barak et al., 2016) and statistical query algorithms (Feldman et al., 2013) have also been shown up to $k = o(\sqrt{n})$.

Secret Leakage PC. We consider a slight generalization of the planted clique problem, where the input graph G comes with some information about the vertex set of the planted clique. This corresponds to the vertices in the k-clique being chosen from some distribution ρ other than the uniform distribution of k-subsets of $[n]$, as formalized in the following definition.

Definition 1 (Secret Leakage PC_ρ) *Given a distribution ρ on k-subsets of $[n]$, let $\mathcal{G}_\rho(n, k, 1/2)$ be the distribution on n-vertex graphs sampled by first sampling $G \sim \mathcal{G}(n, 1/2)$ and $S \sim \rho$ independently and then planting a k-clique on the vertex set S in G. Let $\text{PC}_\rho(n, k, 1/2)$ denote the resulting hypothesis testing problem between $H_0 : G \sim \mathcal{G}(n, 1/2)$ and $H_1 : G \sim \mathcal{G}_\rho(n, k, 1/2)$.*

All of the ρ that we will consider will be uniform over the k-subsets that satisfy some constraint. In the cryptography literature, modifying a problem with a promise of this form is referred to as information leakage about the secret (Kalai and Reyzin, 2019). The hardness of the Learning with Errors (LWE) problem is unconditionally robust to leakage (Dodis et al., 2010; Goldwasser et al., 2010), and it is left as an interesting open problem to show the same is true for PC.

Both PC and PC_ρ fall under the class of general parameter recovery problems where the task is to find P_S generating the observed graph from a family of distributions $\{P_S\}$. In the case of PC, P_S denotes the distribution $\mathcal{G}(n, k, 1/2)$ conditioned on the k-clique being planted on S. Observe that the conditional distributions $\{P_S\}$ are the same in PC and PC_ρ. Secret leakage can be viewed as placing a prior on the parameter S of interest, rather than changing the main average-case part of the problem – the family $\{P_S\}$. When ρ is uniform over a family of k-subsets, secret leakage corresponds to imposing a worst-case constraint on S. In particular, consider the maximum likelihood estimator (MLE) for a general parameter recovery problem given by

$$\hat{S} = \arg \max_{S \in \text{supp}(\rho)} P_S(G)$$

As ρ varies, only the search space of the MLE changes while the objective remains the same. We make the following precise conjecture of the hardness of $\text{PC}_\rho(n, k, 1/2)$ for the distributions ρ we consider. Given a distribution ρ, let $p_\rho(s) = \mathbb{P}_{S,S' \sim \rho^{\otimes 2}}[|S \cap S'| = s]$ be the probability mass function of the size of the intersection of two independent random sets S and S' drawn from ρ.

Conjecture 2 (Secret Leakage Planted Clique (PC_ρ) Conjecture) *Let ρ be one of the distributions on k-subsets of $[n]$ given below in Conjecture 3. Suppose that there is some $p_0 = o_n(1)$, parameter $d = O_n((\log n)^{1+\delta})$ and constant $\delta > 0$ such that $p_\rho(s)$ satisfies the tail bounds*

$$p_\rho(s) \le p_0 \cdot \begin{cases} 2^{-s^2} & \text{if } 1 \le s^2 < d \\ s^{-2d-4} & \text{if } s^2 \ge d \end{cases}$$

Then there is no $\text{poly}(n)$ *time algorithm solving* $\text{PC}_\rho(n, k, 1/2)$.

While this conjecture is only stated for the specific ρ corresponding to the hardness assumptions used in our reductions, we believe it should hold for a wide class of ρ with sufficient symmetry. The motivation for the decay condition on p_ρ in the PC_ρ conjecture is from the *low-degree conjecture* – that low-degree polynomials predict the computational barriers for a broad class of inference problems – which has been shown to match conjectured statistical-computational gaps in a number of problems (Hopkins and Steurer, 2017; Hopkins, 2018; Kunisky et al., 2019; Bandeira et al., 2019). In Section K, we discuss the low-degree conjecture and how it relates to the PC_ρ conjecture, candidate general conditions under which the PC_ρ conjecture holds, and we evidence supporting the PC_ρ conjecture from the failure of low-degree polynomials and SQ lower bounds.

Hardness Conjectures for Specific ρ. In our reductions, we will only need the PC_ρ conjecture for specific ρ, all of which correspond to their own hardness conjectures in natural variants of PC. We now introduce these specific hardness assumptions and briefly outline how each can be produced from an instance of PC_ρ. This is more formally discussed in Section K.1. Let $\mathcal{G}_B(m, n, 1/2)$ denote the distribution on bipartite graphs G with parts of size m and n wherein each edge between the two parts is included independently with probability $1/2$.

- **k-partite PC:** Suppose that k divides n and let E be a partition of $[n]$ into k parts of size n/k. Let $k\text{-PC}_E(n, k, 1/2)$ be $\text{PC}_\rho(n, k, 1/2)$ where ρ is uniformly distributed over all k-sets intersecting each part of E in exactly one element.

- **bipartite PC:** Let $\text{BPC}(m, n, k_m, k_n, 1/2)$ be the problem of testing between $H_0 : G \sim \mathcal{G}_B(m, n, 1/2)$ and H_1 under which G is formed by planting a complete bipartite graph with k_m and k_n vertices in the two parts, respectively, in a graph sampled from $\mathcal{G}_B(m, n, 1/2)$. This problem can be realized as a bipartite subgraph of an instance of PC_ρ.

- **k-part bipartite PC:** Suppose that k_n divides n and let E be a partition of $[n]$ into k_n parts of size n/k_n. Let $k\text{-BPC}_E(m, n, k_m, k_n, 1/2)$ be BPC where the k_n vertices in the part of size n are uniform over all k_n-sets intersecting each part of E in exactly one element, as in the definition of $k\text{-PC}_E$. As with BPC, this problem can be realized as a bipartite subgraph of an instance of PC_ρ, now with additional constraints on ρ to enforce the k-part restriction.

- **k-partite hypergraph PC:** Let k, n and E be as in k-PC. Let $k\text{-HPC}^s_E(n, k, 1/2)$ where $s \ge 3$ be the problem of testing between H_0, under which G is an s-uniform Erdős-Rényi

hypergraph where each hyperedge is included independently with probability $1/2$, and H_1, under which G is first sampled from H_0 and then a k-clique with one vertex chosen uniformly at random from each part of E is planted in G. This problem has a simple correspondence with PC_ρ – there is a specific ρ, stated explicitly in Section K.1, that corresponds to unfolding the adjacency tensor of this hypergraph problem into a matrix.

Since E is always revealed in these problems, it can without loss of generality be taken to be any partition of $[n]$ into k equally-sized parts. We will often simplify notation by dropping the subscript E from the above notation. We conjecture the following computational barriers for these graph problems, each of which matches the decay rate condition on of $p_\rho(s)$ in PC_ρ conjecture, as we will show in Section K.1. Further remarks on Conjectures 2 and 3 can be found in Section B.2.

Conjecture 3 (Specific Hardness Assumptions) *Suppose that m and n are polynomial in one another. Then there is no* $\text{poly}(n)$ *time algorithm solving the following problems:*

1. k-$\text{PC}(n, k, 1/2)$ *when* $k = o(\sqrt{n})$*;*
2. $\text{BPC}(m, n, k_m, k_n, 1/2)$ *when* $k_n = o(\sqrt{n})$ *and* $k_m = o(\sqrt{m})$*;*
3. k-$\text{BPC}(m, n, k_m, k_n, 1/2)$ *when* $k_n = o(\sqrt{n})$ *and* $k_m = o(\sqrt{m})$*; and*
4. k-$\text{HPC}^s(n, k, 1/2)$ *for* $s \geq 3$ *when* $k = o(\sqrt{n})$.

3. Problems and Statistical-Computational Gaps

In this section, we introduce the problems we consider and give informal statements of our main theorems, each of which is a tight computational lower bound implied by one of the assumptions in Conjecture 3. We remark that these lower bounds also follow from planted dense subgraph (PDS) variants of our assumptions or only from the hardness of k-HPC^s, which is the strongest assumption. A much more detailed discussion of our results and further background on the problems we consider can be found in Section B. We also defer stating and discussing our main theorem for one problem – semirandom planted dense subgraph – entirely to Section B.

Statistical Problems and Computational Lower Bounds. Every problem $\mathcal{P}(n, a_1, a_2, \ldots, a_t)$ we consider has a natural parameter n and several other parameters $a_1(n), a_2(n), \ldots, a_t(n)$, which will typically be implicit functions of n. If $\mathcal{P}$ is a hypothesis testing problem with observation X and hypotheses H_0 and H_1, an algorithm $\mathcal{A}$ is deemed to solve $\mathcal{P}$ subject to the constraints $\mathcal{C}$ if it has asymptotic Type I+II error bounded away from 1 when $(n, a_1, a_2, \ldots, a_t) \in \mathcal{C}$ i.e. if $\mathbb{P}_{H_0}[\mathcal{A}(X) = H_1] + \mathbb{P}_{H_1}[\mathcal{A}(X) = H_0] = 1 - \Omega_n(1)$. If $\mathcal{P}$ is an estimation problem with a parameter θ of interest and loss ℓ, then $\mathcal{A}$ solves $\mathcal{P}$ subject to the constraints $\mathcal{C}$ if $\ell(\mathcal{A}(X), \theta) \leq \epsilon$ is true with probability $1 - o_n(1)$ when $(n, a_1, a_2, \ldots, a_t, \epsilon) \in \mathcal{C}$, where $\epsilon = \epsilon(n)$ is a function of n.

We say there is a *computational lower bound* for $\mathcal{P}$ in a parameter regime $\mathcal{C}$ if for any sequence of parameters $\{(n, a_1(n), a_2(n), \ldots, a_t(n))\}_{n=1}^\infty \subseteq \mathcal{C}$ there is another sequence given by $\{(n_i, a_1'(n_i), a_2'(n_i), \ldots, a_t'(n_i))\}_{i=1}^\infty \subseteq \mathcal{C}$ such that $\mathcal{P}(n_i, a_1'(n_i), a_2'(n_i), \ldots, a_t'(n_i))$ cannot be solved in $\text{poly}(n_i)$ time and $\lim_{i\to\infty} \log a_k'(n_i)/\log a_k(n_i) = 1$. In other words, there is a lower bound at $\mathcal{C}$ if, for any sequence s in $\mathcal{C}$, there is another sequence of parameters that cannot be solved in polynomial time and whose growth rate matches the growth rate of a subsequence of s. Thus all of our computational lower bounds are *strong lower bounds* in the sense that rather than show

that a single sequence of parameters is hard, we show that parameter sequences filling out *all possible growth rates* in $\mathcal{C}$ are hard. Throughout the paper, we adopt the standard asymptotic notation $O(\cdot), \Omega(\cdot), o(\cdot), \omega(\cdot)$ and $\Theta(\cdot)$. We let $\tilde{O}(\cdot)$ and analogous variants denote these relations up to $\text{polylog}(n)$ factors, where n is the natural parameter of the problem and will be clear from context.

Robust Sparse Mean Estimation. In sparse mean estimation, the observations $X_1, X_2, \ldots, X_n$ are n independent samples from $\mathcal{N}(\mu, I_d)$ where μ is an unknown k-sparse vector in $\mathbb{R}^d$ of bounded ℓ_2 norm and the task is to estimate μ within an ℓ_2 error of τ. This is a gapless problem, as taking the largest k coordinates of the empirical mean runs in $\text{poly}(d)$ time and achieves the information-theoretically optimal sample complexity of $n = \Theta(k \log d/\tau^2)$.

If an ϵ-fraction of these samples are corrupted arbitrarily by a computationally unbounded adversary, this yields the robust sparse mean estimation problem $\text{RSME}(n, k, d, \tau, \epsilon)$. As discussed in (Li, 2017; Balakrishnan et al., 2017a), the problem is only information-theoretically possible if $\tau = \Omega(\epsilon)$ and estimating at the optimal rate of $\tau = \Theta(\epsilon)$ is only possible with $n = \Theta(k \log d/\epsilon^2)$ samples. However, the best known polynomial-time algorithms require $n = \tilde{\Theta}(k^2 \log d/\epsilon^2)$ samples to estimate μ within $\tau = \Theta(\epsilon\sqrt{\log \epsilon^{-1}})$ in ℓ_2. We give a reduction showing that these polynomial time algorithms are optimal, yielding the first average-case evidence for the k-to-k^2 statistical-computational gap conjectured in Li (2017) and Balakrishnan et al. (2017a). Our reduction also yields the first prediction for the dependence of the computational barrier of RSME on the rate τ.

Theorem 4 (Lower Bounds for RSME) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $\epsilon < 1/2$ is such that (n, ϵ^{-1}) satisfies condition* (T)*, then the k-BPC conjecture implies that there is a computational lower bound for $\text{RSME}(n, k, d, \tau, \epsilon)$ at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

In Theorem 4, (T) denotes a technical condition arising from number-theoretic constraints in our reduction that require that $\epsilon^{-1} = n^{o(1)}$ or $\epsilon^{-1} = \tilde{\Theta}(n^{-1/2t})$ for some positive integer t. As $\epsilon^{-1} = n^{o(1)}$ is the primary regime of interest in the RSME literature, this condition is typically trivial. This condition arises in several of our main theorems and we discuss it in more detail in Section L. We also give an alternate reduction removing it from Theorem 4 in the case where $\epsilon = \tilde{\Theta}(n^{-c})$ for some constant $c \in [0, 1/2]$.

Dense Stochastic Block Models. The stochastic block model (SBM) is the canonical model for community detection and has a long history in the statistics, computer science and statistical physics communities (Abbe, 2017; Moore, 2017). In the k-block SBM, a vertex set of size n is uniformly at random partitioned into k latent communities $C_1, C_2, \ldots, C_k$ each of size n/k and edges are then included in the graph G independently such that intra-community edges appear with probability p while inter-community edges appear with probability $q < p$. The task is to approximately or exactly recover the communities $C_1, C_2, \ldots, C_k$. In the sparse regime when $p = a/n$ and $q = b/n$ for constants a and b, the best known algorithms begin to work at the *Kesten-Stigum threshold* of $\text{SNR} = (a-b)^2/k(a + (k-1)b) > 1$ while the information-theoretic limit is much lower at $\text{SNR} = \Theta(\log k/k)$ (Decelle et al., 2011; Abbe and Sandon, 2018; Banks et al., 2016). The k-block SBM for general edge densities p and q has also been studied extensively under the names graph clustering and graph partitioning (Boppana, 1987; Dyer and Frieze, 1989; Condon and Karp, 2001; McSherry, 2001; Bollobás and Scott, 2004; Chen et al., 2014a; Chen and Xu, 2016). For growing k satisfying $k = O(\sqrt{n})$ and p and q with $p = \Theta(q)$ and $1 - p = \Theta(1 - q)$, the best known $\text{poly}(n)$ time algorithms require $\text{SNR} = (p-q)^2/q(1-q) = \tilde{\Omega}(k^2/n)$, which is an asymptotic extension of

the Kesten-Stigum threshold to general p and q. In contrast, the statistically optimal rate of recovery is again roughly a factor of k lower at $\tilde{\Omega}(k/n)$.

In this work, we show computational lower bounds matching the Kesten-Stigum threshold up to a constant factor in a mean-field analogue of recovering a first community C_1 in the k-SBM, when p and q are bounded away from 0 and 1. Consider a sample G from the k-SBM restricted to the union of the other communities $C_2, \dots, C_k$. This subgraph has average edge density approximately given by $\hat{q} = p/(k-1)+(k-2)q/(k-1)$. We consider the imbalanced SBM problem $\text{ISBM}(n, k, p, q, \hat{q})$, where the task is to recover the community C_1 in the graph G' in which the subgraph on $C_2, \dots, C_k$ is replaced by the corresponding mean-field Erdős-Rényi graph $\mathcal{G}(n - n/k, \hat{q})$. Our main result for ISBM is the following lower bound up to the asymptotic Kesten-Stigum threshold.

Theorem 5 (Lower Bounds for ISBM) *Suppose that (n, k) satisfy condition (T), that k is prime or $k = \omega_n(1)$ and $k = o(n^{1/3})$, and suppose that $q \in (0, 1)$ satisfies $\min\{q, 1 - q\} = \Omega_n(1)$. If $\hat{q} = p/(k - 1) + (k - 2)q/(k - 1)$, then the k-PC conjecture implies that there is a computational lower bound for $\text{ISBM}(n, k, p, q, \hat{q})$ below the Kesten-Stigum threshold of $\frac{(p-q)^2}{q(1-q)} = \tilde{o}(k^2/n)$.*

Testing Hidden Partition Models. We also introduce two testing problems we refer to as the Gaussian and bipartite hidden partition models, which demonstrate the versatility of our reduction technique dense Bernoulli rotations in transforming hidden structure. The task in the bipartite hidden partition model problem is to test for the presence of a planted rK-vertex subgraph, sampled from an r-block stochastic block model, within an n-vertex Erdős-Rényi bipartite graph. Let $\text{BHPM}(n, r, K, P_0, \gamma)$ denote this problem with ambient edge density P_0, edge density $P_0+\gamma$ within the communities in the subgraph and $P_0 - \frac{\gamma}{r-1}$ on the rest of the subgraph. The Gaussian hidden partition model problem $\text{GHPM}(n, r, K, \gamma)$ is a corresponding Gaussian analogue. These problems are formally introduced in Section B.5. As we will show in Section M.2, an empirical variance test succeeds above the threshold $\gamma^2_{\text{comp}} = \tilde{\Theta}(n/rK^2)$ and an exhaustive search succeeds above $\gamma^2_{\text{IT}} = \tilde{\Theta}(1/K)$ in GHPM and BHPM when P_0 is bounded away from 0 and 1. The following theorem states our main lower bounds for these problems, showing that both have a statistical-computational gap and that the empirical variance test is approximately optimal among efficient algorithms.

Theorem 6 (Lower Bounds for GHPM and BHPM) *Suppose that $r^2K^2 = \tilde{\omega}(n)$ and $(\lceil r^2K^2/n \rceil, r)$ satisfies condition (T), suppose r is prime or $r = \omega_n(1)$ and suppose that $P_0 \in (0, 1)$ satisfies $\min\{P_0, 1 - P_0\} = \Omega_n(1)$. Then the k-PC conjecture implies that there is a computational lower bound for each of $\text{GHPM}(n, r, K, \gamma)$ for all levels of signal $\gamma^2 = \tilde{o}(n/rK^2)$. This same lower bound also holds for $\text{BHPM}(n, r, K, P_0, \gamma)$ given the additional condition $n = o(rK^{4/3})$.*

Negatively Correlated Sparse PCA. In the spiked covariance model (Johnstone and Lu, 2004) of sparse PCA, the observations $X = (X_1, X_2, \dots, X_n)$ are n independent samples from either

$$H_0 : X \sim \mathcal{N}(0, I_d)^{\otimes n} \quad \text{and} \quad H_1 : X \sim \mathcal{N}(0, I_d + \theta vv^\top)^{\otimes n}$$

The information-theoretically optimal rate of detection is at the level of signal $\theta = \Theta(\sqrt{k \log d/n})$ (Berthet and Rigollet, 2013b). When $k = o(\sqrt{d})$, the best known polynomial time algorithms for sparse PCA require that $\theta = \Omega(\sqrt{k^2/n})$ and, furthermore, this is implied by the PC conjecture (Berthet and Rigollet, 2013a; Brennan and Bresler, 2019).

In the negatively correlated sparse PCA problem $\text{NEG-SPCA}(n, k, d, \theta)$, the alternative hypothesis is instead given by $H_1 : X \sim \mathcal{N}(0, I_d - \theta vv^\top)^{\otimes n}$. Although the ordinary and negative variants

appear to have the same statistical and computational limits, they are stochastically *very differently structured* problems. A sample from the ordinary spiked covariance model can be expressed as $X_i = \sqrt{\theta} \cdot gv + \mathcal{N}(0, I_d)$ where $g \sim \mathcal{N}(0,1)$ is independent of the $\mathcal{N}(0, I_d)$ term. While this representation is crucial in existing reductions to sparse PCA, negative sparse PCA does not admit a representation of this form, making reductions to the latter problem technically challenging.

Theorem 7 (Lower Bounds for NEG-SPCA) *If $d = \text{poly}(n)$, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, the* BPC *conjecture implies a computational lower bound for* NEG-SPCA(n, k, d, θ) *for $\theta = \tilde{o}(\sqrt{k^2/n})$.*

Our proof of this theorem involves characterizing when a Wishart matrix and its inverse converge in KL divergence, which may be of independent interest. This analysis produces the parameter constraint $k = o(n^{1/6})$ in the theorem above and in our next two theorems, which we believe is an artefact of our techniques and possibly removable. As we will discuss further in Section B.7, conditions of this form *do not affect the tightness* of our lower bounds, but rather only impose a constraint on the level of sparsity k. Our motivation for considering NEG-SPCA is that it has a fundamental connection to the structure of *supervised problems* whereas ordinary sparse PCA does not. In particular, our reduction to NEG-SPCA is a crucial subroutine in reducing to our next two problems.

Unsigned and Mixtures of Sparse Linear Regressions. In learning mixtures of sparse linear regressions (SLR), the observations $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ are n independent sample-label pairs given by $y_i = \langle \beta, X_i \rangle + \eta_i$ where $X_i \sim \mathcal{N}(0, I_d)$, $\eta_i \sim \mathcal{N}(0,1)$ and β is chosen from a mixture distribution ν over a finite set k-sparse vectors $\{\beta_1, \beta_2, \ldots, \beta_L\}$ of bounded ℓ_2 norm (Städler et al., 2010; Wang et al., 2014; Yi and Caramanis, 2015). The task is to estimate the components β_j that are sufficiently likely under ν in ℓ_2 norm i.e. to within an ℓ_2 distance of τ.

We show that a statistical-computational gap emerges for mixtures of SLRs even in the simplest case where there are $L = 2$ components, the mixture distribution ν is known to sample each component with probability $1/2$ and the task is to estimate even just one of the components $\{\beta_1, \beta_2\}$ to within ℓ_2 norm τ. We refer to this simplest setup for learning mixtures of SLRs as MSLR(n, k, d, τ).

Theorem 8 (Lower Bounds for MSLR) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the k-*BPC *conjecture implies that there is a computational lower bound for* MSLR(n, k, d, τ) *at all sample complexities $n = \tilde{o}(k^2/\tau^4)$.*

We will prove this theorem by reducing to the problem of testing between the mixtures of SLRs model when $\beta_1 = -\beta_2$ and a null hypothesis under which y and X are independent. The information-theoretic limit of this testing problem occurs at the sample complexity $n = \tilde{\Theta}(k \log d/\tau^4)$ (Fan et al., 2018), and thus our reduction establishes a k-to-k^2 statistical-computational gap. Our reduction to MSLR also implies a k-to-k^2 gap in a generalized variant of sparse phase retrieval. This provides partial evidence supporting a conjecture in Li and Voroninski (2013), Cai et al. (2016), Wang et al. (2017), Barbier et al. (2019) and Celentano et al. (2020).

Robust Sparse Linear Regression. In SLR, the observations $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ are independent sample-label pairs given by $y_i = \langle \beta, X_i \rangle + \eta_i$ where $X_i \sim \mathcal{N}(0, \Sigma)$, $\eta_i \sim \mathcal{N}(0,1)$ and β is an unknown k-sparse vector with bounded ℓ_2 norm. The task is to estimate β to within ℓ_2 norm τ. The robust SLR problem RSLR$(n, k, d, \tau, \epsilon)$ is obtained when a computationally-unbounded adversary corrupts an arbitrary ϵ-fraction of the observed sample-label pairs. In this work, we consider the simplest case of $\Sigma = I_d$ where ordinary SLR is gapless but robustness seems

to induce a statistical-computational gap. In particular, $n = \Omega(k \log d/\epsilon^2)$ samples are sufficient to solve RSLR when $\tau = \tilde{\Theta}(\epsilon)$ (Liu et al., 2018; Li, 2017), while the best known polynomial-time algorithms require $n = \tilde{\Theta}(k^2 \log d/\epsilon^2)$ samples (Balakrishnan et al., 2017a; Liu et al., 2018). Similar to RSME, robust SLR is only information-theoretically possible if $\tau = \Omega(\epsilon)$ (Gao, 2020). We deduce the following tight computational lower bound for RSLR providing evidence for this conjecture.

Theorem 9 (Lower Bounds for RSLR) *If k, d and n are polynomial in each other, $k = o(n^{1/6})$, $k = o(\sqrt{d})$ and $\epsilon < 1/2$ is such that $\epsilon = \tilde{\Omega}(n^{-1/2})$, then the k-BPC conjecture implies that there is a computational lower bound for $\mathrm{RSLR}(n, k, d, \tau, \epsilon)$ at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

Tensor PCA. In Tensor PCA, the observation is an order s tensor T with dimensions of length n given by $T \sim \theta v^{\otimes s} + \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$, where v has a Rademacher prior and is distributed uniformly over $\{-1, 1\}^n$ (Richard and Montanari, 2014). The task is to recover v within nontrivial ℓ_2 error $o(\sqrt{n})$ and is only information-theoretically possible if $\theta = \tilde{\omega}\left(n^{(1-s)/2}\right)$, while the best known polynomial-time algorithms all require the higher signal strength $\theta = \tilde{\Omega}(n^{-s/4})$ (Richard and Montanari, 2014; Lesieur et al., 2017; Jagannath et al., 2018; Hopkins et al., 2015, 2016; Wein et al., 2019). We show lower bounds against efficient algorithms with a low false positive probability of error in the hypothesis testing formulation of tensor PCA where $T \sim \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$ under H_0 and T is sampled from the tensor PCA distribution described above under H_1. As will be shown in Section N, this implies the same lower bounds hold in the estimation variant of tensor PCA.

Theorem 10 (Lower Bounds for TPCA) *Let n be a parameter and $s \geq 3$ be a constant, then the k-HPC^s conjecture implies a computational lower bound for $\mathrm{TPCA}^s(n, \theta)$ when $\theta = \tilde{o}(n^{-s/4})$ against $\mathrm{poly}(n)$ time algorithms $\mathcal{A}$ solving $\mathrm{TPCA}^s(n, \theta)$ with a low false positive probability of $\mathbb{P}_{H_0}[\mathcal{A}(T) = H_1] = O(n^{-s})$.*

Universality for Learning Sparse Mixtures. So far, the problems we have considered all have either Gaussian or Bernoulli noise distributions. Our final reduction shows that our techniques have implications beyond simple noise distributions, yielding computational lower bounds for the *generalized learning sparse mixtures* problem GLSM. In $\mathrm{GLSM}(n, k, d, \mathcal{U})$ where $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})$, $\mathcal{D}$ is a mixture distribution on $\mathbb{R}$ and the elements of the family $\{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}$ and $\mathcal{Q}$ are distributions on a measurable space satisfying mild conditions outlined in Section F.3. The observations in GLSM are n independent samples $X_1, X_2, \ldots, X_n$ formed as follows:

- for each sample X_i, draw some latent variable $\nu_i \sim \mathcal{D}$ and
- sample $(X_i)_j \sim \mathcal{P}_{\nu_i}$ if $j \in S$ and $(X_i)_j \sim \mathcal{Q}$ otherwise, independently

where S is some unknown subset containing k of the d coordinates. The task in GLSM is to recover S. Given a collection of distributions $\mathcal{U}$ from a wide universality class $\mathrm{UC}(N)$, we define its level of signal $\tau_{\mathcal{U}}$ to be the best asymptotic upper bound on $\left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x)\right|$ for all $\nu \in [-1, 1]$, subject to several additional technical conditions introduced formally in Section B.11. We prove the following computational lower bound for GLSM in terms of $\tau_{\mathcal{U}}$, which generalizes optimal lower bounds for learning sparse mixtures of Gaussians (Azizyan et al., 2013) and sparse PCA.

Theorem 11 (Computational Lower Bounds for GLSM) *Let n, k and d be polynomial in each other and such that $k = o(\sqrt{d})$. Suppose that the collections of distributions $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})$ is in $\mathrm{UC}(n)$. Then the k-BPC conjecture implies a computational lower bound for $\mathrm{GLSM}(n, k, d, \mathcal{U})$ at all sample complexities $n = \tilde{o}\left(\tau_{\mathcal{U}}^{-4}\right)$.*

4. Overview of Techniques

We now give a brief overview of our main techniques and their roles in Figure 1. A much more detailed technical overview of the ideas in this work can be found in Section C.

Dense Bernoulli Rotations. This is a simple reduction primitive that begins with a random vector of independent entries in $\{0,1\}^n$ and one unknown bit i with an elevated mean and produces a vector from $\mathcal{N}(A_i, I_m)$, approximately in total variation without knowing i. The prescribed mean vectors $A_1, A_2, \ldots, A_n \in \mathbb{R}^m$ can be any set satisfying that the largest singular value of the matrix A with columns A_i is at most $O(1/\sqrt{\log n})$. This yields a general structure-transforming primitive that will be used throughout our reductions. Each such use will consist of many local applications of dense Bernoulli rotations to sub-blocks of the adjacency matrix of the input PC_ρ instance.

Design Matrices and Tensors. In each of our applications of dense Bernoulli rotations, the key technical obstacle is designing a set of vectors $A_1, A_2, \ldots, A_n$ that simultaneously matches the combinatorial structure of the target distribution and does not degrade the level of signal in the input. We introduce several families of matrices based on the incidence geometry of finite fields achieving these two objectives in the problems we reduce to. We also give an involved design tensor construction based on this geometry and linear functions in $\mathbb{F}_r$ for GHPM and BHPM, as well as an alternative random matrix construction for our reductions to RSME and RSLR.

Decomposing Linear Regression and Label Generation. Our reductions to MSLR and RSLR are motivated by the observation that X conditioned on y in a single sample from either of these problems can be expressed as the independent sum of: (1) the output of our reduction to RSME, and (2) a sample from NEG-SPCA. Furthermore, when MSLR and RSLR are near their computational barriers, so are both of the instances in this sum. Our reductions use this decomposition along with a technique for generating a sample (X, y) given $X|y = y'$ where y' is from a fixed binary set.

Producing Negative Correlations and Inverse Wishart Matrices. As RSME can be mapped to with dense Bernoulli rotations, to produce MSLR and RSLR, a cloning trick shows that it suffices to reduce to NEG-SPCA. Our main idea is to produce samples $X_1, X_2, \ldots, X_n \in \mathbb{R}^m$ from ordinary sparse PCA by applying a reduction of Brennan and Bresler (2019) and then to output the columns of $\hat{\Sigma}^{-1/2}R$ where $\hat{\Sigma}$ is a rescaling of the empirical covariance matrix of the X_i's and R is an independent $m \times n$ matrix sampled from Haar measure on the Stiefel manifold. To prove the correctness of this reduction, we characterize when a Wishart matrix and its inverse converge in KL divergence, which may be of independent interest.

Completing Tensors from Hypergraphs and Tensor PCA. In order to apply dense Bernoulli rotations in our reduction to TPCA, it is crucial to complete missing diagonal entries in the adjacency tensor of the input k-HPC^s instance to produce an instance of a planted subtensor problem. This involves an iterative cloning procedure, causing our reduction to require multiple queries to a tensor PCA blackbox, as opposed to a typical reduction in total variation which requires a single query.

Symmetric 3-ary Rejection Kernels and Universality. Our final reduction technique is a variant of the rejection kernels introduced in Brennan et al. (2018) and Brennan et al. (2019a) designed to show tight lower bounds for learning sparse mixtures where ordinary rejection kernels fail to. This technique is the key step in our reduction to GLSM.

Acknowledgments

We are greatly indebted to Jerry Li for introducing the conjectured statistical-computational gap for robust sparse mean estimation and for discussions that helped lead to this work. We thank Ilias Diakonikolas for pointing out the statistical query model construction in Diakonikolas et al. (2017). We thank the anonymous reviewers for helpful feedback that greatly improved the exposition. We also thank Frederic Koehler, Sam Hopkins, Philippe Rigollet, Enric Boix-Adserà, Dheeraj Nagaraj, Rares-Darius Buhai, Alex Wein, Ilias Zadik, Dylan Foster and Austin Stromme for inspiring discussions on related topics. This work was supported in part by MIT-IBM Watson AI Lab and NSF CAREER award CCF-1940205.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

Emmanuel Abbe and Colin Sandon. Achieving the ks threshold in the general stochastic block model with linearized acyclic belief propagation. In *Advances in Neural Information Processing Systems*, pages 1334–1342, 2016.

Emmanuel Abbe and Colin Sandon. Proof of the achievability conjectures for the general stochastic block model. *Communications on Pure and Applied Mathematics*, 71(7):1334–1406, 2018.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2015.

Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 793–802. IEEE, 2008.

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

Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 496–505. ACM, 2007.

Brendan PW Ames. Guaranteed clustering and biclustering via semidefinite programming. *Mathematical Programming*, 147(1-2):429–465, 2014.

Brendan PW Ames and Stephen A Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical programming*, 129(1):69–89, 2011.

Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15 (1):2239–2312, 2014.

Benny Applebaum, Boaz Barak, and David Xiao. On basing lower-bounds for learning on worst-case assumptions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 211–220. IEEE, 2008.

Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458. ACM, 2014.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147, 2013.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Efficient sparse clustering of high-dimensional non-spherical gaussian mixtures. In *Artificial Intelligence and Statistics*, pages 37–45, 2015.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. pages 169–212, 2017a.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017b.

Afonso S Bandeira and Ramon Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

Afonso S Bandeira, Amelia Perry, and Alexander S Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *arXiv preprint arXiv:1803.11132*, 2018.

Afonso S Bandeira, Dmitriy Kunisky, and Alexander S Wein. Computational hardness of certifying bounds on constrained pca problems. *arXiv preprint arXiv:1902.07324*, 2019.

Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416, 2016.

Boaz Barak. *The Complexity of Public-Key Cryptography*, pages 45–77. Springer International Publishing, Cham, 2017. ISBN 978-3-319-57048-8. doi: 10.1007/978-3-319-57048-8_2. URL https://doi.org/10.1007/978-3-319-57048-8_2.

Boaz Barak, Samuel B Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 428–437. IEEE, 2016.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *arXiv preprint arXiv:1711.05424*, 2017.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *arXiv preprint arXiv:1808.00921*, 2018.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, pages 1046–1066, 2013a.

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

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995.

Andrej Bogdanov and Luca Trevisan. Average-case complexity. *Foundations and Trends® in Theoretical Computer Science*, 2(1):1–106, 2006a.

Andrej Bogdanov and Luca Trevisan. On worst-case to average-case reductions for np problems. *SIAM Journal on Computing*, 36(4):1119–1159, 2006b.

Béla Bollobás and Alex D Scott. Max cut for random graphs with a planted partition. *Combinatorics, Probability and Computing*, 13(4-5):451–474, 2004.

Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

Ravi B Boppana. Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*, pages 280–285. IEEE, 1987.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

Matthew Brennan and Guy Bresler. Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness. In *Conference on Learning Theory*, pages 469–470, 2019.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *COLT*, pages 48–166, 2018.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Universality of computational lower bounds for submatrix detection. In *Conference on Learning Theory*, pages 417–468, 2019a.

Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *arXiv preprint arXiv:1910.14167*, 2019b.

Sébastien Bubeck and Shirshendu Ganguly. Entropic clt and phase transition in high-dimensional wishart matrices. *International Mathematics Research Notices*, 2018(2):588–606, 2016.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures and Algorithms*, 2016.

Thang Nguyen Bui, Soma Chaudhuri, Frank Thomson Leighton, and Michael Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7(2):171–191, 1987.

Cristina Butucea and Yuri I Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

T. Tony Cai and Yihong Wu. Statistical and computational limits for sparse matrix detection. *arXiv preprint arXiv:1801.00518*, 2018.

T Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*, 2015a.

Tony Cai, Zongming Ma, and Yihong Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3-4):781–815, 2015b.

Francesco Caltagirone, Marc Lelarge, and Léo Miolane. Recovering asymmetric communities in the stochastic block model. *IEEE Transactions on Network Science and Engineering*, 5(3):237–246, 2018.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Utkan Onur Candogan and Venkat Chandrasekaran. Finding planted subgraphs with few eigenvalues using the schur–horn relaxation. *SIAM Journal on Optimization*, 28(1):735–759, 2018.

Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. *arXiv preprint arXiv:2002.12903*, 2020.

Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pages 1040–1048, 2013.

Yao-ban Chan and Peter Hall. Using evidence of mixed populations to select variables for clustering very high-dimensional data. *Journal of the American Statistical Association*, 105(490):798–809, 2010.

Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pages 35–1, 2012.

Wei-Kuo Chen. Phase transition in the spiked random tensor with rademacher prior. *The Annals of Statistics*, 47(5):2734–2756, 2019.

Wei-Kuo Chen, Madeline Handschy, and Gilad Lerman. Phase transition in random tensors with multiple spikes. *arXiv preprint arXiv:1809.06790*, 2018.

Wei-Kuo Chen, David Gamarnik, Dmitry Panchenko, and Mustazee Rahman. Suboptimality of local algorithms for a class of max-cut problems. *The Annals of Probability*, 47(3):1587–1618, 2019.

Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

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*, 17(27):1–57, 2016.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in neural information processing systems*, pages 2204–2212, 2012.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014a.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pages 560–604, 2014b.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017a.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Transactions on Information Theory*, 64(3):1738–1766, 2017b.

Didier Chételat and Martin T Wells. The middle-scale asymptotics of wishart matrices. *The Annals of Statistics*, 47(5):2639–2670, 2019.

Hyonho Chun and Sündüz Keles. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, 2009.

Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

Anne Condon and Richard M Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*, 18(2):116–140, 2001.

Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. pages 441–448, 2014.

Roee David and Uriel Feige. On the effect of randomness on planted 3-coloring models. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 77–90. ACM, 2016.

Richard D De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8 (3):227–245, 1989.

Scott E Decatur, Oded Goldreich, and Dana Ron. Computational sample complexity. *SIAM Journal on Computing*, 29(3):854–879, 2000.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. *Combinatorics, Probability and Computing*, 23(1):29–49, 2014.

Yash Deshpande and Andrea Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, 2015a.

Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. pages 523–562, 2015b.

Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*, 2015.

Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664. IEEE, 2016.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. Society for Industrial and Applied Mathematics, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606, 2019a.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019b.

Yevgeniy Dodis, Shafi Goldwasser, Yael Tauman Kalai, Chris Peikert, and Vinod Vaikuntanathan. Public-key encryption schemes with auxiliary inputs. In *Theory of Cryptography Conference*, pages 361–381. Springer, 2010.

Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *Advances in Neural Information Processing Systems*, pages 6065–6075, 2019.

Martin E. Dyer and Alan M. Frieze. The solution of some random np-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

Ronen Eldan and Dan Mikulincer. Information and dimensionality of anisotropic random geometric graphs. *arXiv preprint arXiv:1609.02490*, 2016.

Jianqing Fan, Han Liu, Zhaoran Wang, and Zhuoran Yang. Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval. *arXiv preprint arXiv:1808.06996*, 2018.

Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Uriel Feige. Relations between average case complexity and approximation complexity. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 534–543. ACM, 2002.

Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.

Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures and Algorithms*, 16(2):195–208, 2000.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, pages 189–204. Discrete Mathematics and Theoretical Computer Science, 2010.

Joan Feigenbaum and Lance Fortnow. Random-self-reducibility of complete sets. *SIAM Journal on Computing*, 22(5):994–1005, 1993.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 655–664. ACM, 2013.

Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. pages 77–86, 2015.

Zoltán Füredi and János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. *The Annals of Probability*, 45(4):2353–2376, 2017.

David Gamarnik and Ilias Zadik. High dimensional regression with binary coefficients. estimating squared error and a phase transtition. In *Conference on Learning Theory*, pages 948–953, 2017.

David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property. *arXiv preprint arXiv:1904.07174*, 2019.

Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

Chao Gao, Zongming Ma, and Harrison H Zhou. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Oded Goldreich. Notes on Levin's theory of average-case complexity. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 233–247. Springer, 2011.

Shafi Goldwasser, Yael Kalai, Chris Peikert, and Vinod Vaikuntanathan. Robustness of the learning with errors assumption. In *Innovations in Computer Science*, pages 230–240, 2010.

Dima Grigoriev. Linear lower bound on degrees of positivstellensatz calculus proofs for the parity. *Theoretical Computer Science*, 259(1-2):613–622, 2001.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016a.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937, 2016b.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Information limits for recovering a hidden community. pages 1894–1898, 2016c.

Bruce E Hajek, Yihong Wu, and Jiaming Xu. Computational lower bounds for community detection on random graphs. In *COLT*, pages 899–928, 2015.

Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pages 9136–9146, 2018.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Justin Holmgren and Alexander Wein. Counterexamples to the low-degree conjecture. *arXiv preprint arXiv:2004.08454*, 2020.

Samuel B Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

Samuel B Hopkins and Jerry Li. How hard is robust mean estimation? *arXiv preprint arXiv:1903.07870*, 2019.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 379–390. IEEE, 2017.

Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *COLT*, pages 956–1006, 2015.

Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. *Proceedings of the fifty-eighth IEEE Foundations of Computer Science*, pages 720–731, 2017.

Samuel B Hopkins, Pravesh Kothari, Aaron Henry Potechin, Prasad Raghavendra, and Tselil Schramm. On the integrality gap of degree-4 sum of squares for planted clique. *ACM Transactions on Algorithms (TALG)*, 14(3):1–31, 2018.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

Peter J Huber. *Robust statistics*. Springer, 2011.

Aukosh Jagannath, Patrick Lopatto, and Leo Miolane. Statistical thresholds for tensor pca. *arXiv preprint arXiv:1812.03403*, 2018.

Tiefeng Jiang and Danning Li. Approximation of rectangular beta-laguerre ensembles and large deviations. *Journal of Theoretical Probability*, 28(3):804–847, 2015.

Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *Unpublished manuscript*, 2004.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Yael Tauman Kalai and Leonid Reyzin. A survey of leakage-resilient cryptography. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 727–794. 2019.

Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. In *Advances in Neural Information Processing Systems*, pages 7423–7432, 2019.

Richard M Karp. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Mathematics of operations research*, 2(3):209–224, 1977.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430, 2018.

Pascal Koiran and Anastasios Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE Transactions on Information Theory*, 60(8):4999–5006, 2014.

Alexandra Kolla, Konstantin Makarychev, and Yury Makarychev. How to play unique games against a semi-random adversary: Study of semi-random models of unique games. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 443–452. IEEE, 2011.

Pravesh K Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer. Sum of squares lower bounds for refuting any csp. *arXiv preprint arXiv:1701.04521*, 2017.

Akshay Krishnamurthy, Arya Mazumdar, Andrew McGregor, and Soumyabrata Pal. Sample complexity of learning mixture of sparse linear regressions. In *Advances in Neural Information Processing Systems*, pages 10531–10540, 2019.

Florent Krzakała, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104(25):10318–10323, 2007.

L Kučera. Expected behavior of graph coloring algorithms. In *International Conference on Fundamentals of Computation Theory*, pages 447–451. Springer, 1977.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 680–687. IEEE, 2015.

Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering. In *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*, pages 601–608. IEEE, 2016.

Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 511–515. IEEE, 2017.

Leonid A Levin. Average case complete problems. *SIAM Journal on Computing*, 15(1):285–286, 1986.

Jerry Li. Robust sparse estimation tasks in high dimensions. *arXiv preprint arXiv:1702.05860*, 2017.

Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pages 1125–1144, 2018.

Nathan Linial. Locality in distributed graph algorithms. *SIAM Journal on Computing*, 21(1):193–201, 1992.

Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust estimation of sparse models via trimmed hard thresholding. *arXiv preprint arXiv:1901.08237*, 2019.

Linyuan Lu and Xing Peng. Spectra of edge-independent random graphs. *The Electronic Journal of Combinatorics*, 20(4):P27, 2013.

Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse pca. In *Advances in Neural Information Processing Systems*, pages 1612–1620, 2015.

Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

A Majumdar. Image compression by sparse pca coding in curvelet domain. *Signal, image and video processing*, 3(1):27–34, 2009.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 367–384. ACM, 2012.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Correlation clustering with noisy partial information. In *Conference on Learning Theory*, pages 1321–1342, 2015.

Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün. Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing*, 26(1-2):303–324, 2016.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, London, 1979.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

Claire Mathieu and Warren Schudy. Correlation clustering with noisy input. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 712–728. Society for Industrial and Applied Mathematics, 2010.

Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68, 2011.

Cathy Maugis, Gilles Celeux, and Marie-Laure Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709, 2009.

Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares lower bounds for planted clique. pages 87–96, 2015.

Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting sum-of-squares lower bounds: Degree-2 to degree-4. *arXiv preprint arXiv:1911.01411*, 2019.

Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841. ACM, 2016.

Marco Mondelli and Andrea Montanari. On the connection between learning two-layers neural networks and tensor decomposition. *arXiv preprint arXiv:1802.07301*, 2018a.

Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450, 2018b.

Andrea Montanari. Finding one community in a sparse graph. *Journal of Statistical Physics*, 161 (2):273–299, 2015.

Cristopher Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

Raj Rao Nadakuditi and Mark EJ Newman. Graph spectra and the detectability of community structure in networks. *Physical review letters*, 108(18):188701, 2012.

Joe Neeman and Praneeth Netrapalli. Non-reconstructability in the stochastic block model. *arXiv preprint arXiv:1404.6304*, 2014.

Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164, 2007.

Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8(1):1–34, 2009.

Amelia Perry and Alexander S Wein. A semidefinite program for unbalanced multisection in the stochastic block model. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 64–67. IEEE, 2017.

Amelia Perry, Alexander S Wein, and Afonso S Bandeira. Statistical limits of spiked tensor models. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 230–264. Institut Henri Poincaré, 2020.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Richard E Quandt and James B Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American statistical Association*, 73(364):730–738, 1978.

Miklós Z Rácz and Jacob Richey. A smooth transition from wishart to goe. *Journal of Theoretical Probability*, 32(2):898–906, 2019.

Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM, 2020.

Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. *arXiv preprint arXiv:1807.11419*, 6, 2018.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

Alexander A Razborov and Steven Rudich. Natural proofs. *Journal of Computer and System Sciences*, 55(1):24–35, 1997.

Reinhold Remmert. *Classical topics in complex function theory*, volume 172. Springer Science & Business Media, 2013.

Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Typology of phase transitions in Bayesian inference problems. *Physical Review E*, 99(4):042109, 2019.

Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Valentina Ros, Gerard Ben Arous, Giulio Biroli, and Chiara Cammarota. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Physical Review X*, 9(1):011003, 2019.

Benjamin Rossman. On the constant-depth complexity of k-clique. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 721–730. ACM, 2008.

Benjamin Rossman. The monotone complexity of k-clique on random graphs. *SIAM Journal on Computing*, 43(1):256–279, 2014.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

Philip Schniter and Sundeep Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2014.

Rocco A Servedio. Computational sample complexity and attribute-efficient learning. In *Proceedings of the thirty-first annual ACM symposium on Theory of Computing*, pages 701–710, 1999.

Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. ℓ_1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Nicolas Verzelen and Ery Arias-Castro. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.

Nicolas Verzelen and Ery Arias-Castro. Detection and feature selection in sparse mixture models. *The Annals of Statistics*, 45(5):1920–1950, 2017.

Aravindan Vijayaraghavan and Pranjal Awasthi. Clustering semi-random mixtures of gaussians. In *International Conference on Machine Learning*, pages 5055–5064, 2018.

Van H Vu. Spectral norm of random matrices. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 423–430. ACM, 2005.

Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. Online object tracking with sparse prototypes. *IEEE transactions on image processing*, 22(1):314–325, 2013.

Gang Wang, Liang Zhang, Georgios B Giannakis, Mehmet Akçakaya, and Jie Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491, 2017.

Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of rip certification. In *Advances in Neural Information Processing Systems*, pages 3819–3827, 2016a.

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

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

Michel Wedel and Wayne S DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of classification*, 12(1):21–55, 1995.

Alexander S Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor pca. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1446–1468. IEEE, 2019.

John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.

Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.

Xinyang Yi and Constantine Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

Dong Yin, Ramtin Pedarsani, Yudong Chen, and Kannan Ramchandran. Learning mixtures of sparse linear regressions using sparse graph codes. *IEEE Transactions on Information Theory*, 65(3):1430–1451, 2018.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *arXiv preprint arXiv:1703.02724*, 2017.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *COLT*, pages 921–948, 2014.

Hong-Tu Zhu and Heping Zhang. Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):3–16, 2004.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Contents

Part I
Detailed Background and Overview

In this part, we expand upon the previous sections to give more detailed background on average-case reductions, an in-depth discussion of the problems we consider and our main theorems and a detailed overview of our technical contributions and reduction primitives. We remark that we include some redundancy for clarity of exposition.

Appendix A. Statistical-Computational Gaps and Average-case Reductions

In this section, we give a more detailed description of the prior work on statistical-computational gaps and average-case reductions. We also outline four general criteria for a reduction between statistical problems to show strong computational lower bounds, and discuss the technical obstacles arising in devising these reductions.

A.1. Deferred Background from Section 1

Failure of Classes of Algorithms. In the last few years, there have been many exciting developments in the line of research analyzing when powerful classes of algorithms fail to solve average-case problems. A breakthrough work of Barak et al. (2016) developed the general technique of pseudocalibration for showing SOS lower bounds, and used this method to prove tight lower bounds for planted clique (PC). In Hopkins (2018), pseudocalibration motivated a general conjecture on the optimality of low-degree polynomials for hypothesis testing that has been used to provide evidence for a number of additional gaps (Hopkins and Steurer, 2017; Kunisky et al., 2019; Bandeira et al., 2019). There have also been many other recent SOS lower bounds (Grigoriev, 2001; Deshpande and Montanari, 2015b; Ma and Wigderson, 2015; Meka et al., 2015; Kothari et al., 2017; Hopkins et al., 2018; Raghavendra et al., 2018; Hopkins et al., 2017; Mohanty et al., 2019). Other classes of algorithms for which there has been progress in a similar vein include statistical query algorithms (Feldman et al., 2013, 2015; Diakonikolas et al., 2017, 2019b), classes of circuits (Razborov and Rudich, 1997; Rossman, 2008, 2014), local algorithms (Gamarnik and Sudan, 2017; Linial, 1992) and message-passing algorithms (Zdeborová and Krzakala, 2016; Lesieur et al., 2015, 2016; Krzakała et al., 2007; Ricci-Tersenghi et al., 2019; Bandeira et al., 2018). Another line of work has aimed to provide evidence for computational limits by establishing properties of the energy landscape of solutions that are barriers to natural optimization-based approaches (Achlioptas and Coja-Oghlan, 2008; Gamarnik and Zadik, 2017; Ben Arous et al., 2017, 2018; Ros et al., 2019; Chen et al., 2019; Gamarnik and Zadik, 2019).

Background on Average-Case Reductions. While there has been success analyzing when these classes of algorithms fail to solve average-case problems, progress towards a traditional reduction-based approach to their computational complexity has been more limited. This is because reductions between average-case problems are more constrained and overall very different from reductions between worst-case problems. Average-case combinatorial problems have been studied in computer science since the 1970's (Karp, 1977; Kučera, 1977). In the 1980's, Levin introduced his theory of average-case complexity (Levin, 1986), formalizing the notion of an average-case reduction and obtaining abstract completeness results. Since then, average-case complexity has been studied

extensively in cryptography and complexity theory. A survey of this literature can be found in Bogdanov and Trevisan (2006a) and Goldreich (2011). As discussed in (Barak, 2017) and (Goldreich, 2011), average-case reductions are notoriously delicate and there is a lack of available techniques. Although technically difficult to obtain, average-case reductions have a number of advantages over other approaches. Aside from the advantage of being future-proof against new classes of algorithms, showing that a problem of interest is hard by reducing from PC effectively *subsumes* hardness for classes of algorithms known to fail on PC and thus gives stronger evidence for hardness. Reductions preserving gaps also directly relate phenomena across problems and reveal insights into how parameters, hidden structures and noise models correspond to one another.

Worst-case reductions are only concerned with transforming the *hidden structure* in one problem to another. For example, a worst-case reduction from 3-SAT to k-INDEPENDENT-SET needs to ensure that the hidden structure of a satisfiable 3-SAT formula is mapped to a graph with an independent set of size k, and that an unsatisfiable formula is not. Average-case reductions need to not only transform the structure in one problem to that of another, but also precisely map between the *natural distributions* associated with problems. In the case of the example above, all classical worst-case reductions use gadgets that map random 3-SAT formulas to a very unnatural distribution on graphs. Average-case problems in statistical inference are also fundamentally *parameterized*, with parameter regimes in which the problem is information-theoretically impossible, possible but conjecturally computationally hard and computationally easy. To establish the strongest possible lower bounds, reductions need to exactly fill out one of these three parameter regimes – the one in which the problem is conjectured to be computationally hard. These subtleties that arise in devising average-case reductions will be discussed further in the next section.

A.2. Desiderata for Average-Case Reductions

As discussed in the previous section, average-case reductions are delicate and more constrained than their worst-case counterparts. In designing average-case reductions between problems in statistical inference, the essential challenge is to reduce to instances that are *hard up to the conjectured computational barrier*, without destroying the *naturalness* of the distribution over instances. Dissecting this objective further yields four general criteria for a reduction between the problems $\mathcal{P}$ and $\mathcal{P}'$ to be deemed to show strong computational lower bounds for $\mathcal{P}'$. These objectives are to varying degrees at odds with one another, which is what makes devising reductions a challenging task. To illustrate these concepts, our running example will be our reduction from PC_ρ to robust sparse linear regression (SLR). Some parts of this discussion are slightly simplified for clarity. The following are our four criteria.

1. **Aesthetics:** If $\mathcal{P}$ and $\mathcal{P}'$ each have a specific canonical distribution then a reduction must faithfully map these distributions to one another. In our example, this corresponds to mapping the independent 0-1 edge indicators in a random graph to noisy Gaussian samples of the form $y = \langle \beta, X \rangle + \mathcal{N}(0,1)$ with $X \sim \mathcal{N}(0, I_d)$ and where an ϵ-fraction are corrupted.

2. **Mapping Between Different Structures:** A reduction must simultaneously map all possible latent signals of $\mathcal{P}$ to that of $\mathcal{P}'$. In our example, this corresponds to mapping each possible clique position in PC_ρ to a specific mixture over the hidden vector β. A reduction in this case would also need to map between possibly very differently structured data, e.g., in robust SLR the dependence of (X, y) on β is intricate and the ϵ-fraction of corrupted samples also

produces latent structure across samples. These are both very different than the planted signal plus noise form of the clique in PC_ρ.

3. **Tightness to Algorithms:** A reduction showing computational lower bounds that are tight against what efficient algorithms can achieve needs to map the conjectured computational limits of $\mathcal{P}$ to those of $\mathcal{P}'$. In our example, PC_ρ in general has a conjectured limit depending on ρ, which may for instance be at $K = o(\sqrt{N})$ when the clique is of size K in a graph with N vertices. In contrast, robust SLR has the conjectured limit at $n = \tilde{o}(k^2\epsilon^2/\tau^4)$, where τ is the ℓ_2 error to which we wish to estimate β, k is the sparsity of β and n is the number of samples.

4. **Strong Lower Bounds for Parameterized Problems:** In order to show that a certain constraint $\mathcal{C}$ *defines* the computational limit of $\mathcal{P}'$ through this reduction, we need the reduction to fill out the possible parameter sequences within $\mathcal{C}$. For example, to show that $n = \tilde{o}(k^2\epsilon^2/\tau^4)$ truly captures the correct dependence in our computational lower bound for robust SLR, it does not suffice to produce a single sequence of points $(n, k, d, \tau, \epsilon)$ for which this is true, or even a one parameter curve. There are four parameters in the conjectured limit and a reduction showing that this is the correct dependence needs to fill out any possible combination of growth rates in these parameters permitted by $n = \tilde{o}(k^2\epsilon^2/\tau^4)$. The fact that the initial problem $\mathcal{P}$ has a conjectured limit depending on only two parameters can make achieving this criterion challenging.

We remark that the third criterion requires that reductions are *information preserving* in the sense that they do not degrade the underlying level signal used by optimal efficient algorithms. This necessitates that the amount of additional randomness introduced in reductions to achieve aesthetic requirements is negligible. The fourth criterion arises from the fact that statistical problems are generally described by a tuple of parameters and are therefore actually an entire family of problems. A full characterization of the computational feasibility of a problem therefore requires addressing all possible scalings of the parameters.

All of the reductions carried out in this paper satisfy all four desiderata. Several of the initial reductions from PC in the literature met most but not all of these criteria. For example, the reductions in Berthet and Rigollet (2013a) and Wang et al. (2016b) to sparse PCA map to a distribution in a distributionally robust formulation of the problem as opposed to the canonical Gaussian formulation in the spiked covariance model. Similarly Cai et al. (2015a) reduces to a distributionally robust formulation of submatrix localization. The reduction in Gao et al. (2017) only shows tight computational lower bounds for sparse PCA at a particular point in the parameter space when $\theta = \tilde{\Theta}(1)$ and $n = \tilde{\Theta}(k^2)$. However, a number of reductions in the literature have successfully met all of these four criteria (Ma and Wu, 2015; Hajek et al., 2015; Zhang and Xia, 2017; Brennan et al., 2018; Brennan and Bresler, 2019; Brennan et al., 2019a).

We remark that it can be much easier to only satisfy some of these desiderata – in particular, many natural reduction ideas meet a subset of these four criteria but fail to show nontrivial computational lower bounds. For instance, it is often straightforward to construct a reduction that degrades the level of signal. The simple reduction that begins with PC and randomly subsamples edges with probability $n^{-\alpha}$ yields an instance of planted dense subgraph with the correct distributional aesthetics. However, this reduction fails to be tight to algorithms and furthermore fails to show any meaningful tradeoff between the size of the planted dense subgraph and the sparsity of the graph.

Another natural reduction to robust sparse mean estimation first maps from PC to Gaussian biclustering using one of the reductions in Ma and Wu (2015), Brennan et al. (2018) or Brennan et al. (2019a), computes the sum v of all of the rows of this matrix, then uses Gaussian cloning as in Brennan et al. (2018) to produce n weak copies of v and finally outputs these copies with an an ϵ-fraction corrupted. This reduction can be verified to produce a valid instance of robust sparse mean estimation in its canonical Gaussian formulation, but fails to show any nontrivial hardness above its information-theoretic limit. Conceptually, this is because the reduction is generating the ϵ-fraction of the corruptions itself. On applying a robust sparse mean estimation blackbox to solve PC, the reduction could just as easily have revealed which samples it corrupted. This would allow the blackbox to only have to solve sparse mean estimation, which has no statistical-computational gap. In general, a reduction showing tight computational lower bounds cannot generate a non-negligible amount randomness that produces the hardness of the target problem. Instead, this ϵ-fraction must come from the hidden clique in the input PC instance. In Section C.8, we discuss how our reductions obliviously encode cliques into the hidden structures in the problems we consider.

We also remark that many problems that appear to be similar from the perspective of designing efficient algorithms can be quite different to reduce to. This arises from differences in their underlying stochastic models that efficient algorithms do not have to make use of. For example, although ordinary sparse PCA and sparse PCA with a negative spike can be solved by the same efficient algorithms, the former has a signal plus noise decomposition while the latter does not and has negatively correlated as opposed to positively correlated planted entries. We will see that these subtle differences are significant in designing reductions.

Appendix B. Detailed Overview of Problems and Main Results

In this section, we give detailed background on each of the problems introduced in Section 3 and discuss the conditions in our main theorems in more depth. For convenience and clarity, we repeat all of our main theorems and some of the discussion from Section 3.

B.1. Remarks on Our Computational Lower Bounds

In this section, we make some further remarks on general features of the computational lower bounds in our main theorems. Each of our computational lower bounds for estimation problems will be established through a reduction to a hypothesis testing problem which then implies the desired lower bound. The exact formulations for these intermediate hypothesis testing problems can be found in Section E.3 and how they also imply lower bounds for estimation and recovery variants of our problems is discussed in Section P. Throughout this work, we will use the terms detection and hypothesis testing interchangeably. All of our reductions are to the canonical simplest average-case formulations of the problems we consider. For example, all k-sparse unit vectors in our lower bounds are binary and in $\{0, 1/\sqrt{k}\}^d$, and the rank-1 component in our lower bound for tensor PCA is sampled from a Rademacher prior. Our reductions are all also to the canonical simple vs. simple hypothesis testing formulation for each of our problems and, as discussed in Brennan et al. (2018), this yields strong computational lower bounds, is often technically more difficulty and crucially allows reductions to naturally be composed with one another.

B.2. Remarks on the PC$_\rho$ Conjecture

In this section, we make several further remarks on the PC$_\rho$ conjecture deferred from Section 2.

The motivation for the decay condition on p_ρ in the PC$_\rho$ conjecture is from low-degree polynomials, which we show in Section K.2 fail to solve PC$_\rho$ subject to this condition. The *low-degree conjecture* – that low-degree polynomials predict the computational barriers for a broad class of inference problems – has been shown to match conjectured statistical-computational gaps in a number of problems (Hopkins and Steurer, 2017; Hopkins, 2018; Kunisky et al., 2019; Bandeira et al., 2019). We discuss this conjecture, the technical conditions arising in its formalizations and how these relate to PC$_\rho$ in Section K.2. Specifically, we discuss the importance of symmetry and the requirement on d in generalizing Conjecture 2 to further ρ. In contrast to low-degree polynomials, because the SQ model only concerns problems with samples, it seems ill-suited to accurately predict the computational barriers in PC$_\rho$ for every ρ. However, in Section K.3, we show SQ lower bounds supporting the PC$_\rho$ conjecture for specific ρ related to our hardness assumptions. We also remark that the distribution p_ρ is an overlap distribution, which has been linked to statistical-computational gaps using techniques from statistical physics (Zdeborová and Krzakala, 2016).

While we only need the specific hardness assumptions in Conjecture 3 to deduce our computational lower bounds, secret leakage can be viewed as a way to conceptually unify these different assumptions. The ρ corresponding to the problems in Conjecture 3 all seem to avoid revealing enough information about S to give rise to new polynomial time algorithms to solve PC$_\rho$. In particular, spectral algorithms consistently seem to match our conjectured computational limits for PC$_\rho$ for the different ρ we consider. From an entropy viewpoint, the k-partite assumption common to these variants of PC$_\rho$ only reveals a very small amount of information about the location of the clique. In particular, both the uniform distribution over k-subsets and over k-subsets respecting a given partition E have $(1+o(1))k\log_2 n$ bits of entropy.

We also remark that the PC$_\rho$ conjecture, as stated, implies the thresholds in Conjecture 3 up to arbitrarily small polynomial factors i.e. where the thresholds are $k = O(n^{1/2-\epsilon})$, $k_n = O(n^{1/2-\epsilon})$ and $k_m = O(m^{1/2-\epsilon})$ for arbitrarily small $\epsilon > 0$. As we will discuss in K.2, the low-degree conjecture also supports the stronger thresholds in Conjecture 3. We also note that our reductions continue to show tight hardness up to arbitrarily small polynomial factors even under these weaker assumptions. As mentioned in Section 1.1, our hardness assumption for k-HPC^s is the strongest of those in Conjecture 3. Specifically, in Section K.1 we give simple reductions showing that (4) in Conjecture 3 implies (1), (2) and (3).

We remark that the discussion in Section 2 also applies when planted clique is replaced with the planted dense subgraph (PDS) problem. In the PDS variant of a PC problem, instead of planting a k-clique in $\mathcal{G}(n,1/2)$, a dense subgraph $\mathcal{G}(k,p)$ is planted in $\mathcal{G}(n,q)$ where $p > q$. We conjecture that all of the hardness assumptions remain true for PDS with constant edge densities $0 < q < p \leq 1$. Note that PC is an instance of PDS with $p = 1$ and $q = 1/2$. All of the reductions beginning with PC$_\rho$ in this work will also yield reductions beginning from secret leakage planted dense subgraph problems PDS$_\rho$. In particular, they will continue to apply with a small loss in the amount of signal when $q = 1/2$ and $p = 1/2 + n^{-\epsilon}$ for a small constant $\epsilon > 0$. As discussed in Brennan and Bresler (2019), PDS conjecturally has no quasipolynomial time algorithms in this regime and thus our reductions would transfer lower bounds above polynomial time. In this parameter regime, the barriers of PDS also appear to be similar to those of detection in the sparsely spiked Wigner model, which also conjecturally has no quasipolynomial time algorithms (Hopkins et al.,

2017). Throughout this work, we will denote the PDS variants of the problems introduced above by k-PDS(n, k, p, q), BPDS(m, n, k_m, k_n, p, q), k-BPDS(m, n, k_m, k_n, p, q) and k-HPDS$^s(n, k, p, q)$.

B.3. Robust Sparse Mean Estimation

The study of robust estimation began with Huber's contamination model (Huber, 1992, 1965) and observations of Tukey (Tukey, 1975). Classical robust estimators have typically either been computationally intractable or heuristic (Huber, 2011; Tukey, 1975; Yatracos, 1985). Recent breakthrough works (Diakonikolas et al., 2016; Lai et al., 2016) gave the first efficient algorithms for high-dimensional robust estimation, which sparked an active line of research into robust algorithms for other high-dimensional problems (Awasthi et al., 2014; Li, 2017; Balakrishnan et al., 2017a; Charikar et al., 2017; Diakonikolas et al., 2018; Klivans et al., 2018; Diakonikolas et al., 2019b; Hopkins and Li, 2019; Dong et al., 2019). The most canonical high-dimensional robust estimation problem is robust sparse mean estimation, which has an intriguing statistical-computational gap induced by robustness.

In sparse mean estimation, the observations $X_1, X_2, \ldots, X_n$ are n independent samples from $\mathcal{N}(\mu, I_d)$ where μ is an unknown k-sparse vector in $\mathbb{R}^d$ of bounded ℓ_2 norm and the task is to estimate μ within an ℓ_2 error of τ. This is a gapless problem, as taking the largest k coordinates of the empirical mean runs in poly(d) time and achieves the information-theoretically optimal sample complexity of $n = \Theta(k \log d/\tau^2)$.

If an ϵ-fraction of these samples are corrupted arbitrarily by an adversary, this yields the robust sparse mean estimation problem RSME$(n, k, d, \tau, \epsilon)$. As discussed in Li (2017) and Balakrishnan et al. (2017a), for $\|\mu - \mu'\|_2$ sufficiently small, it holds that $d_{\mathrm{TV}}(\mathcal{N}(\mu, I_d), \mathcal{N}(\mu', I_d)) = \Theta(\|\mu - \mu'\|_2)$. Furthermore, an ϵ-corrupted set of samples can simulate distributions within $O(\epsilon)$ total variation from $\mathcal{N}(\mu, I_d)$. Therefore ϵ-corruption can simulate $\mathcal{N}(\mu', I_d)$ if $\|\mu' - \mu\|_2 = O(\epsilon)$ and it is impossible to estimate μ with ℓ_2 distance less than this $O(\epsilon)$. This implies that the minimax rate of estimation for μ is $O(\epsilon)$, even for very large n. As shown in Li (2017) and Balakrishnan et al. (2017a), the information-theoretic threshold for estimating at this rate in the ϵ-corrupted model remains at $n = \Theta(k \log d/\epsilon^2)$ samples. However, the best known polynomial-time algorithms from Li (2017) and Balakrishnan et al. (2017a) require $n = \tilde{\Theta}(k^2 \log d/\epsilon^2)$ samples to estimate μ within $\tau = \epsilon\sqrt{\log \epsilon^{-1}}$ in ℓ_2. In Sections I.1 and L.1, we give a reduction showing that these polynomial time algorithms are optimal, yielding the first average-case evidence for the k-to-k^2 statistical-computational gap conjectured in Li (2017) and Balakrishnan et al. (2017a). Our reduction applies to more general rates τ and obtains the following tradeoff.

Theorem 4 (Lower Bounds for RSME) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $\epsilon < 1/2$ is such that (n, ϵ^{-1}) satisfies condition* (T)*, then the k-BPC conjecture implies that there is a computational lower bound for* RSME$(n, k, d, \tau, \epsilon)$ *at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

For example, taking $\epsilon = 1/3$ and $\tau = \tilde{O}(1)$ shows that there is a k-to-k^2 gap between the information-theoretically optimal sample complexity of $n = \tilde{\Theta}(k)$ and the computational lower bound of $n = \tilde{o}(k^2)$. Note that taking $\tau = O(\epsilon)$ in Theorem 4 recovers exactly the tradeoff in Li (2017) and Balakrishnan et al. (2017a), with the dependence on ϵ. Our reduction to RSME is based on dense Bernoulli rotations and constructions of combinatorial design matrices based on incidence geometry in $\mathbb{F}_r^t$, as is further discussed in Sections 4 and G.

In Theorem 4, (T) denotes a technical condition arising from number-theoretic constraints in our reduction that require that $\epsilon^{-1} = n^{o(1)}$ or $\epsilon^{-1} = \tilde{\Theta}(n^{-1/2t})$ for some positive integer t. As $\epsilon^{-1} = n^{o(1)}$ is the primary regime of interest in the RSME literature, this condition is typically trivial. We discuss the condition (T) in more detail in Section L and give an alternate reduction removing it from Theorem 4 in the case where $\epsilon = \tilde{\Theta}(n^{-c})$ for some constant $c \in [0, 1/2]$.

Our result also holds in the stronger Huber's contamination model where an ϵ-fraction of the n samples are chosen at random and replaced with i.i.d. samples from another distribution $\mathcal{D}$. The prior work of Diakonikolas et al. (2017) showed that SQ algorithms require $n = \tilde{\Omega}(k^2)$ samples to solve RSME, establishing the conjectured k-to-k^2 gap in the SQ model. However, our work is the first to make a precise prediction of the computational barrier in RSME as a function of both ϵ and τ. As will be discussed in Section I.1, our reduction from k-PC maps to the instance of RSME under the adversary introduced in Diakonikolas et al. (2017).

B.4. Dense Stochastic Block Models

The stochastic block model (SBM) is the canonical model for community detection, having independently emerged in the machine learning and statistics (Holland et al., 1983), computer science (Bui et al., 1987; Dyer and Frieze, 1989; Boppana, 1987), statistical physics (Decelle et al., 2011) and mathematics communities (Bollobás et al., 2007). It has been the subject of a long line of research, which has recently been surveyed in Abbe (2017) and Moore (2017). In the k-block SBM, a vertex set of size n is uniformly at random partitioned into k latent communities $C_1, C_2, \ldots, C_k$ each of size n/k and edges are then included in the graph G independently such that intra-community edges appear with probability p while inter-community edges appear with probability $q < p$. The exact recovery problem entails finding $C_1, C_2, \ldots, C_k$ and the weak recovery problem, also known as community detection, entails outputting nontrivial estimates $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$ with $|C_i \cap \hat{C}_i| \geq (1 + \Omega(1))n/k$.

Community detection in the SBM is often considered in the sparse regime, where $p = a/n$ and $q = b/n$. In Decelle et al. (2011), non-rigorous arguments from statistical physics were used to form the precise conjecture that weak recovery begins to be possible in poly(n) time exactly at the *Kesten-Stigum* threshold $\text{SNR} = (a-b)^2/k(a + (k-1)b) > 1$. When $k = 2$, the algorithmic side of this conjecture was confirmed with methods based on belief propagation (Mossel et al., 2018), spectral methods and non-backtracking walks (Massoulié, 2014; Bordenave et al., 2015), and it was shown to be information-theoretically impossible to solve weak recovery below the Kesten-Stigum threshold in (Mossel et al., 2015; Deshpande et al., 2015). The algorithmic side of this conjecture for general k was subsequently resolved with approximate acyclic belief propagation in (Abbe and Sandon, 2015, 2016, 2018) and has also been shown using low-degree polynomials, tensor decomposition and color coding (Hopkins and Steurer, 2017). A statistical-computational gap is conjectured to already arise at $k = 4$ (Abbe and Sandon, 2018) and the information-theoretic limit for community detection has been shown to occur for large k at $\text{SNR} = \Theta(\log k/k)$, which is much lower than the Kesten-Stigum threshold (Banks et al., 2016). Rigorous evidence for this statistical-computational gap has been much more elusive and has only been shown for low-degree polynomials (Hopkins and Steurer, 2017) and variants of belief propagation. Another related line of work has exactly characterized the thresholds for exact recovery in the regime $p, q = \Theta(\log n/n)$ when $k = 2$ (Abbe et al., 2015; Hajek et al., 2016a,b).

The k-block SBM for general edge densities p and q has also been studied extensively under the names graph clustering and graph partitioning in the statistics and computer science communities. A long line of work has developed algorithms recovering the latent communities in this regime, including a wide range of spectral and convex programming techniques (Boppana, 1987; Dyer and Frieze, 1989; Condon and Karp, 2001; McSherry, 2001; Bollobás and Scott, 2004; Coja-Oghlan, 2010; Rohe et al., 2011; Chaudhuri et al., 2012; Nadakuditi and Newman, 2012; Chen et al., 2012; Ames, 2014; Anandkumar et al., 2014; Chen et al., 2014a; Chen and Xu, 2016). A comparison and survey of these results can be found in Chen et al. (2014a). As discussed in Chen and Xu (2016), for growing k satisfying $k = O(\sqrt{n})$ and p and q with $p = \Theta(q)$ and $1 - p = \Theta(1 - q)$, the best known poly(n) time algorithms all only work above

$$\frac{(p-q)^2}{q(1-q)} \gtrsim \frac{k^2}{n}$$

which is an asymptotic extension of the Kesten-Stigum threshold to general p and q. In contrast, the statistically optimal rate of recovery is again roughly a factor of k lower at $\tilde{\Omega}(k/n)$. Furthermore, up to $\log n$ factors, the Kesten-Stigum threshold is both when efficient exact recovery algorithms begin to work and where the best efficient weak recovery algorithms are conjectured to fail (Chen and Xu, 2016).

In this work, we show computational lower bounds matching the Kesten-Stigum threshold up to a constant factor in a mean-field analogue of recovering a first community C_1 in the k-SBM, where p and q are bounded away from zero and one. Consider a sample G from the k-SBM restricted to the union of the other communities $C_2, \dots, C_k$. This subgraph has average edge density approximately given by $\hat{q} = (p-q)\cdot(k-1)\cdot(n/k)^2\cdot(n-n/k)^{-2}+q = (k-1)^{-1}\cdot p+(1-(k-1)^{-1})\cdot q$. Now consider the task of recovering the community C_1 in the graph G' in which the subgraph on $C_2, \dots, C_k$ is replaced by the corresponding mean-field Erdős-Rényi graph $\mathcal{G}(n - n/k, \hat{q})$. Formally, let G' be the graph formed by first choosing C_1 at random and sampling edges as follows:

- include edges within C_1 with probability $P_{11} = p$;
- include edges between C_1 and $[n]\backslash C_1$ with probability $P_{12} = q$; and
- includes edges within $[n]\backslash C_1$ with probability P_{22} where $P_{22} = (k-1)^{-1} \cdot p + (1 - (k-1)^{-1}) \cdot q$.

We refer to this model as the imbalanced SBM and let ISBM$(n, k, P_{11}, P_{12}, P_{22})$ denote the problem of testing between this model and Erdős-Rényi graphs of the form $\mathcal{G}(n, P_0)$. As we will discuss in Section E.3, lower bounds for this formulation also imply lower bounds for weakly and exactly recovering C_1. We remark that under our notation for ISBM, the hidden community C_1 has size n/k and k is the number of communities in the analogous k-block SBM described above.

As we will discuss in Section M.1, ISBM can also be viewed as a model of single community detection with uniformly calibrated expected degrees. Note that the expected degree of a vertex in C_1 is $nP_{22} - p$ and the expected degree of a vertex in $C_1\backslash[n]$ is $(n-1)P_{22}$, which differ by at most 1. Similar models with two imbalanced communities and calibrated expected degrees have appeared previously in Neeman and Netrapalli (2014), Verzelen and Arias-Castro (2015), Perry and Wein (2017) and Caltagirone et al. (2018). As will be discussed in Section B.6, the simpler planted dense subgraph model of single community recovery has a detection threshold that differs

from the Kesten-Stigum threshold, even though the Kesten-Stigum threshold is conjectured to be the barrier for recovering the planted dense subgraph. This is because non-uniformity in expected degrees gives rise to simple edge-counting tests that do not lead to algorithms for recovering the planted subgraph. Our main result for ISBM is the following lower bound up to the asymptotic Kesten-Stigum threshold.

Theorem 5 (Lower Bounds for ISBM) *Suppose that (n, k) satisfy condition* (T)*, that k is prime or $k = \omega_n(1)$ and $k = o(n^{1/3})$, and suppose that $q \in (0, 1)$ satisfies $\min\{q, 1-q\} = \Omega_n(1)$. If $P_{22} = (k-1)^{-1} \cdot p + (1 - (k-1)^{-1}) \cdot q$, then the k-PC conjecture implies that there is a computational lower bound for $\mathrm{ISBM}(n, k, p, q, P_{22})$ at all levels of signal below the Kesten-Stigum threshold of $\frac{(p-q)^2}{q(1-q)} = \tilde{o}(k^2/n)$.*

This directly provides evidence for the conjecture that $(p-q)^2/q(1-q) = \tilde{\Theta}(k^2/n)$ defines the computational barrier for community recovery in general k-SBMs made in Chen and Xu (2016). While the statistical-computational gaps in PC and k-SBM are the two most prominent conjectured gaps in average-case problems over graphs, they are very different from an algorithmic perspective and evidence for computational lower bounds up to the Kesten-Stigum threshold has remained elusive. Our reduction yields a first step towards understanding the relationship between these gaps.

B.5. Testing Hidden Partition Models

We also introduce two testing problems we refer to as the Gaussian and bipartite hidden partition models. We give a reduction and algorithms that show these problems have a statistical-computational gap, and we tightly characterize their computational barriers based on the k-PC conjecture. The main motivation for introducing these problems is to demonstrate the versatility of our reduction technique dense Bernoulli rotations in transforming hidden structure. A description of dense Bernoulli rotations and the construction of a key design tensor used in our reduction can be found in Section G.

The task in the bipartite hidden partition model problem is to test for the presence of a planted rK-vertex subgraph, sampled from an r-block stochastic block model, in an n-vertex random bipartite graph. The Gaussian hidden partition model problem is a corresponding Gaussian analogue. These are both multi-community variants of the subgraph stochastic block model considered in Brennan et al. (2018), which corresponds to the setting in which $r = 2$. The multi-community nature of the planted subgraph yields a more intricate hidden structure, and the additional free parameter r yields a more complicated computational barrier. The work of Chen and Xu (2016) considered the related task of recovering the communities in the Gaussian and bipartite hidden partition models. We remark that conjectured computational limits for this recovery task differ from the detection limits we consider.

Formally, our hidden partition problems are defined as follows. Let $C = (C_1, C_2, \ldots, C_r)$ and $D = (D_1, D_2, \ldots, D_r)$ are chosen independently and uniformly at random from the set of all sequences of length r consisting of disjoint K-subsets of $[n]$. Consider the random matrix M sampled by first sampling C and D and then sampling

$$M_{ij} \sim \begin{cases} \mathcal{N}(\gamma, 1) & \text{if } i \in C_h \text{ and } j \in D_h \text{ for some } h \in [r] \\ \mathcal{N}\left(-\frac{\gamma}{r-1}, 1\right) & \text{if } i \in C_{h_1} \text{ and } j \in D_{h_2} \text{ where } h_1 \neq h_2 \\ \mathcal{N}(0, 1) & \text{otherwise} \end{cases}$$

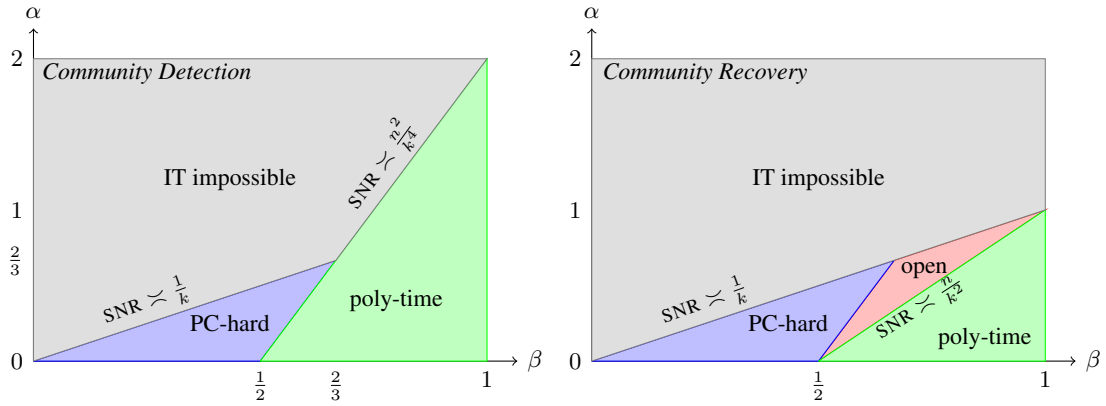

Figure 2: Prior computational and statistical barriers in the detection and recovery of a single hidden community from the PC conjecture (Hajek et al., 2015; Brennan et al., 2018, 2019a). The axes are parameterized by α and β where $\text{SNR} = \frac{(P_1-P_0)^2}{P_0(1-P_0)} = \tilde{\Theta}(n^{-\alpha})$ and $k = \tilde{\Theta}(n^{\beta})$. The red region is conjectured to be hard but no PC reductions showing this are known.

independently for each $1 \le i, j \le n$. The problem $\text{GHPM}(n, r, K, \gamma)$ is to test between $H_0 : M \sim \mathcal{N}(0,1)^{\otimes n \times n}$ and an alternative hypothesis H_1 under which M is sampled as outlined above. The problem $\text{BHPM}(n, r, K, P_0, \gamma)$ is a bipartite graph analogue of this problem with ambient edge density P_0, edge density $P_0 + \gamma$ within the communities in the subgraph and $P_0 - \frac{\gamma}{r-1}$ on the rest of the subgraph.

As we will show in Section M.2, an empirical variance test succeeds above the threshold $\gamma^2_{\text{comp}} = \tilde{\Theta}(n/rK^2)$ and an exhaustive search succeeds above $\gamma^2_{\text{IT}} = \tilde{\Theta}(1/K)$ in GHPM and BHPM where P_0 is bounded away from 0 and 1. Thus our main lower bounds for these two problems confirm that this empirical variance test is approximately optimal among efficient algorithms and that both problems have a statistical-computational gap assuming the k-PC conjecture.

Theorem 6 (Lower Bounds for GHPM and BHPM) *Suppose that $r^2K^2 = \tilde{\omega}(n)$ and $(\lceil r^2K^2/n \rceil, r)$ satisfies condition* (T)*, suppose r is prime or $r = \omega_n(1)$ and suppose that $P_0 \in (0,1)$ satisfies $\min\{P_0, 1-P_0\} = \Omega_n(1)$. Then the k-PC conjecture implies that there is a computational lower bound for each of $\text{GHPM}(n, r, K, \gamma)$ for all levels of signal $\gamma^2 = \tilde{o}(n/rK^2)$. This same lower bound also holds for $\text{BHPM}(n, r, K, P_0, \gamma)$ given the additional condition $n = o(rK^{4/3})$.*

We also remark that the empirical variance and exhaustive search tests along with our lower bound do not support the existence of a statistical-computational gap in the case when the subgraph is the entire graph with $n = rK$, which is our main motivation for considering this subgraph variant. We remark that a number of the technical conditions in the theorem such as condition (T) and $n = o(rK^{4/3})$ are trivial in the parameter regime where the number of communities is not very large with $r = n^{o(1)}$ and when the total size of the hidden communities is large with $rK = \tilde{\Theta}(n^c)$ where $c > 3/4$. In this regime, these problems have a nontrivial statistical-computational gap that our result tightly characterizes.

B.6. Semirandom Planted Dense Subgraph and the Recovery Conjecture

In the planted dense subgraph model of single community recovery, the observation is a sample from $\mathcal{G}(n, k, P_1, P_0)$ which is formed by planting a random subgraph on k vertices from $\mathcal{G}(k, P_1)$ inside a copy of $\mathcal{G}(n, P_0)$, where $P_1 > P_0$ are allowed to vary with n and satisfy that $P_1 = O(P_0)$. Detection and recovery of the hidden community in this model have been studied extensively (Arias-Castro and Verzelen, 2014; Butucea and Ingster, 2013; Verzelen and Arias-Castro, 2015; Hajek et al., 2015; Chen and Xu, 2016; Hajek et al., 2016c; Montanari, 2015; Candogan and Chandrasekaran, 2018) and this model has emerged as a canonical example of a problem with a detection-recovery computational gap. While it is possible to efficiently detect the presence of a hidden subgraph of size $k = \tilde{\Omega}(\sqrt{n})$ if $(P_1 - P_0)^2/P_0(1 - P_0) = \tilde{\Omega}(n^2/k^4)$, the best known polynomial time algorithms to *recover* the subgraph require a higher signal at the Kesten-Stigum threshold of $(P_1 - P_0)^2/P_0(1 - P_0) = \tilde{\Omega}(n/k^2)$.

In each of Hajek et al. (2015), Brennan et al. (2018) and Brennan et al. (2019a), it has been conjectured that the recovery problem is hard below this threshold of $\tilde{\Theta}(n/k^2)$. This PDS Recovery Conjecture was even used in Brennan et al. (2018) as a hardness assumption to show detection-recovery gaps in other problems including biased sparse PCA and Gaussian biclustering. A line of work has tightly established the conjectured detection threshold through reductions from the PC conjecture (Hajek et al., 2015; Brennan et al., 2018, 2019a), while the recovery threshold has remained elusive. Planted clique maps naturally to the detection threshold in this model, so it seems unlikely that the PC conjecture could also yield lower bounds at the tighter recovery threshold, given that recovery and detection are known to be equivalent for PC Alon et al. (2007). These prior lower bounds and the conjectured detection-recovery gap in PDS are depicted in Figure 2.

We show that the k-PC conjecture implies the PDS Recovery Conjecture for *semirandom* community recovery in the regime where $q = \Theta(1)$. Semirandom adversaries provide an alternate notion of robustness against constrained modifications that heuristically appear to increase the signal strength (Blum and Spencer, 1995). Algorithms and lower bounds in semirandom problems have been studied for a number of problems, including the stochastic block model (Feige and Kilian, 2001; Moitra et al., 2016), planted clique (Feige and Krauthgamer, 2000), unique games (Kolla et al., 2011), correlation clustering (Mathieu and Schudy, 2010; Makarychev et al., 2015), graph partitioning (Makarychev et al., 2012), 3-coloring (David and Feige, 2016) and clustering mixtures of Gaussians (Vijayaraghavan and Awasthi, 2018). Formally we consider the problem SEMI-CR(n, k, P_1, P_0) where a semirandom adversary is allowed to remove edges outside of the planted subgraph from a graph sampled from $\mathcal{G}(n, k, P_1, P_0)$. The task is to test between this model and an Erdős-Rényi graph $\mathcal{G}(n, P_0)$ similarly perturbed by a semirandom adversary. As we will discuss in Section E.3, lower bounds for this formulation extend to approximately recovering the community under a semirandom adversary. In Section M.3, we prove the following theorem – that the computational barrier in the detection problem shifts to the recovery threshold in SEMI-CR.

Theorem 12 (Lower Bounds for SEMI-CR) *If k and n are polynomial in each other with $k = \Omega(\sqrt{n})$ and $0 < P_0 < P_1 \leq 1$ where $\min\{P_0, 1 - P_0\} = \Omega(1)$, then the k-PC conjecture implies that there is a computational lower bound for* SEMI-CR(n, k, P_1, P_0) *at $\frac{(P_1-P_0)^2}{P_0(1-P_0)} = \tilde{o}(n/k^2)$.*

A related reference is the reduction in Cai et al. (2015a), which proves a detection-recovery gap in the context of sub-Gaussian submatrix localization based on the hardness of finding a planted k-clique in a random $n/2$-regular graph. The relationship between our lower bound and that of

Cai et al. (2015a) is discussed in more detail in Section M.3. From an algorithmic perspective, the convexified maximum likelihood algorithm from Chen and Xu (2016) complements our lower bound – a simple monotonicity argument shows that it continues to solve the community recovery problem above the Kesten-Stigum threshold under a semirandom adversary.

B.7. Negatively Correlated Sparse Principal Component Analysis

In sparse principal component analysis (PCA), the observations $X_1, X_2, \ldots, X_n$ are n independent samples from $\mathcal{N}(0, \Sigma)$ where the eigenvector v corresponding to the largest eigenvalue of Σ is k-sparse, and the task is to estimate v in ℓ_2 norm or find its support. Sparse PCA has many applications ranging from online visual tracking (Wang et al., 2013) and image compression (Majumdar, 2009) to gene expression analysis (Zou et al., 2006; Chun and Keles, 2009; Parkhomenko et al., 2009; Chan and Hall, 2010). Showing lower bounds for sparse PCA can be reduced to analyzing detection in the spiked covariance model (Johnstone and Lu, 2004), which has hypotheses

$$H_0 : X \sim \mathcal{N}(0, I_d)^{\otimes n} \quad \text{and} \quad H_1 : X \sim \mathcal{N}(0, I_d + \theta vv^\top)^{\otimes n}$$

Here, H_1 is the composite hypothesis where $v \in \mathbb{R}^d$ is unknown and allowed to vary over all k-sparse unit vectors. The information-theoretically optimal rate of detection is at the level of signal $\theta = \Theta(\sqrt{k \log d/n})$ (Berthet and Rigollet, 2013b; Cai et al., 2015b; Wang et al., 2016b). However, when $k = o(\sqrt{d})$, the best known polynomial time algorithms for sparse PCA require that $\theta = \Omega(\sqrt{k^2/n})$. Since the seminal paper of Berthet and Rigollet (2013a) initiated the study of statistical-computational gaps through the PC conjecture, this k-to-k^2 gap for sparse PCA has been shown to follow from the PC conjecture in a sequence of papers (Berthet and Rigollet, 2013b,a; Wang et al., 2016b; Gao et al., 2017; Brennan et al., 2018; Brennan and Bresler, 2019).

In negatively correlated sparse PCA, the eigenvector v of interest instead corresponds to the *smallest eigenvalue* of Σ. Negative sparse PCA can similarly be formulated as a hypothesis testing problem NEG-SPCA(n, k, d, θ), where the alternative hypothesis is instead given by $H_1 : X \sim \mathcal{N}(0, I_d - \theta vv^\top)^{\otimes n}$. Similar algorithms as in ordinary sparse PCA continue to work in the negative setting – the information-theoretic limit of the problem remains at $\theta = \Theta(\sqrt{k \log d/n})$ and the best known efficient algorithms still require $\theta = \Omega(\sqrt{k^2/n})$. However, negative sparse PCA is stochastically a *very differently structured* problem than ordinary sparse PCA. A sample from the ordinary spiked covariance model can be expressed as

$$X_i = \sqrt{\theta} \cdot gv + \mathcal{N}(0, I_d)$$

where $g \sim \mathcal{N}(0, 1)$ is independent of the $\mathcal{N}(0, I_d)$ term. This signal plus noise representation is a common feature in many high-dimensional statistical models and is crucially used in the reductions showing hardness for sparse PCA in Berthet and Rigollet (2013b), Berthet and Rigollet (2013a), Wang et al. (2016b), Gao et al. (2017), Brennan et al. (2018), and Brennan and Bresler (2019). Negative sparse PCA does not admit a representation of this form, making it an atypical planted problem and different from ordinary sparse PCA, despite the deceiving similarity between their optimal algorithms. The lack of this representation makes reducing to Negative sparse PCA technically challenging. Negatively spiked PCA was also recently related to the hardness of finding approximate ground states in the Sherrington-Kirkpatrick model (Bandeira et al., 2019). However, ordinary PCA does not seem to share this connection. In Section H, we give a reduction obtaining the following computational lower bound for NEG-SPCA from the BPC conjecture.

Theorem 7 (Lower Bounds for NEG-SPCA) *If $d = \text{poly}(n)$, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the* BPC *conjecture implies a computational lower bound for* NEG-SPCA(n, k, d, θ) *at all levels of signal $\theta = \tilde{o}(\sqrt{k^2/n})$.*

We deduce this theorem and discuss its conditions in detail in Section L.2. A key step in our reduction to NEG-SPCA involves randomly rotating the positive semidefinite square root of the inverse of an empirical covariance matrix. In analyzing this step, we prove a novel convergence result in random matrix theory, which may be of independent interest. Specifically, we characterize when a Wishart matrix and its inverse converge in KL divergence. This is where the parameter constraint $k = o(n^{1/6})$ in the theorem above arises. We believe that this is an artefact of our techniques and extending the theorem to hold without this condition is an interesting open problem. A similar condition arose in the strong lower bounds of Brennan and Bresler (2019). We remark that conditions of this form *do not affect the tightness* of our lower bounds, but rather only impose a constraint on the level of sparsity k. More precisely, for each fixed level of sparsity $k = \tilde{\Theta}(n^{\alpha})$, there is conjectured statistical-computational gap in θ between the information-theoretic barrier of $\theta = \Theta(\sqrt{k \log d/n})$ and computational barrier of $\theta = \tilde{o}(\sqrt{k^2/n})$. Our reduction tightly establishes this gap for all $\alpha \in (0, 1/6]$. Our main motivation for considering NEG-SPCA is that it seems to have a fundamental connection to the structure of *supervised problems* where ordinary sparse PCA does not. In particular, our reduction to NEG-SPCA is a crucial subroutine in reducing to mixtures of sparse linear regressions and robust sparse linear regression. This is discussed further in Sections 4, H and I.

B.8. Unsigned and Mixtures of Sparse Linear Regressions

In learning mixtures of sparse linear regressions (SLR), the task is to learn L sparse linear functions capturing the relationship between features and response variables in heterogeneous samples from L different sparse regression problems. Formally, the observations $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ are n independent sample-label pairs given by $y_i = \langle \beta, X_i \rangle + \eta_i$ where $X_i \sim \mathcal{N}(0, I_d)$, $\eta_i \sim \mathcal{N}(0, 1)$ and β is chosen from a mixture distribution ν over a finite set k-sparse vectors $\{\beta_1, \beta_2, \ldots, \beta_L\}$ of bounded ℓ_2 norm. The task is to estimate the components β_j that are sufficiently likely under ν in ℓ_2 norm i.e. to within an ℓ_2 distance of τ.

Mixtures of linear regressions, also known as the hierarchical mixtures of experts model in the machine learning community (Jordan and Jacobs, 1994), was first introduced in Quandt and Ramsey (1978) and has been studied extensively in the past few decades (De Veaux, 1989; Wedel and DeSarbo, 1995; McLachlan and Peel, 2004; Zhu and Zhang, 2004; Faria and Soromenho, 2010). Recent work on mixtures of linear regressions has focussed on efficient algorithms with finite-sample guarantees (Chaganty and Liang, 2013; Chen et al., 2014b; Yi et al., 2014; Balakrishnan et al., 2017b; Chen et al., 2017b; Li and Liang, 2018). The high-dimensional setting of mixtures of SLRs was first considered in Städler et al. (2010), which proved an oracle inequality for an ℓ_1-regularization approach, and variants of the EM algorithm for mixtures of SLRs were analyzed in Wang et al. (2014) and Yi and Caramanis (2015). Recent work has also studied a different setting for mixtures of SLRs where the covariates X_i can be designed by the learner (Yin et al., 2018; Krishnamurthy et al., 2019).

We show that a statistical-computational gap emerges for mixtures of SLRs even in the simplest case where there are $L = 2$ components, the mixture distribution ν is known to sample each component with probability $1/2$ and the task is to estimate even just one of the components $\{\beta_1, \beta_2\}$ to

within ℓ_2 norm τ. We refer to this simplest setup for learning mixtures of SLRs as MSLR(n, k, d, τ). The following computational lower bound is deduced in Section L.3 and is a consequence of the reduction in Section I.

Theorem 8 (Lower Bounds for MSLR) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the k-BPC conjecture implies that there is a computational lower bound for* MSLR(n, k, d, τ) *at all sample complexities $n = \tilde{o}(k^2/\tau^4)$.*

As we will discuss in Section E.3, we will prove this theorem by reducing to the problem of testing between the mixtures of SLRs model when $\beta_1 = -\beta_2$ and a null hypothesis under which y and X are independent. A closely related work (Fan et al., 2018) studies a nearly identical testing problem in the statistical query model. They tightly characterize the information-theoretic limit of this problem, showing that it occurs at the sample complexity $n = \tilde{\Theta}(k \log d/\tau^4)$. Therefore our reduction establishes a k-to-k^2 statistical-computational gap in this model of learning mixtures of SLRs. In Fan et al. (2018), it is also shown that efficient algorithms in the statistical query model suffer from this same k-to-k^2 gap.

Our reduction to the hypothesis testing formulation of MSLR above is easily seen to imply that the same computational lower bound holds for an unsigned variant USLR(n, k, d, τ) of SLR, where the n observations $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ now of the form $y_i = |\langle \beta, X_i \rangle + \eta_i|$ for a fixed unknown β. Note that by the symmetry of $\mathcal{N}(0, 1)$, y_i is equidistributed to $||\langle \beta, X_i \rangle| + \eta_i|$ and thus is a noisy observation of $|\langle \beta, X_i \rangle|$. In general, noisy observations of the phaseless modulus $|\langle \beta, X_i \rangle|$ from some conditional link distribution $\mathbb{P}(\cdot \,|\, |\langle \beta, X_i \rangle|)$ yields a general instance of phase retrieval (Mondelli and Montanari, 2018b; Celentano et al., 2020). As observed in Fan et al. (2018), the problem USLR is close to the canonical formulation of sparse phase retrieval (SPR) where $\mathbb{P}(\cdot \,|\, |\langle \beta, X_i \rangle|)$ is $\mathcal{N}(|\langle \beta, X_i \rangle|^2, \sigma^2)$, which has been studied extensively and has a conjectured k-to-k^2 statistical-computational gap (Li and Voroninski, 2013; Schniter and Rangan, 2014; Candes et al., 2015; Cai et al., 2016; Wang et al., 2017; Hand et al., 2018; Barbier et al., 2019; Celentano et al., 2020). Our lower bounds provide partial evidence for this conjecture and it is an interesting open problem to give a reduction to the canonical formulation of SPR and other sparse GLMs through average-case reductions.

The reduction to MSLR showing Theorem 8 in Section I is our capstone reduction. It showcases a wide range of our techniques including dense Bernoulli rotations, constructions of combinatorial design matrices from $\mathbb{F}_r^t$, our reduction to NEG-SPCA and its connection to random matrix theory, and an additional technique of combining instances of different unsupervised problems into a supervised problem. We give an overview of these techniques in Section 4. Furthermore, MSLR is a very differently structured problem from any of our variants of PC and it is surprising that the tight statistical-computational gap for MSLR can be derived from their hardness. We remark that our lower bounds for MSLR inherit the technical condition that $k = o(n^{1/6})$ from our reduction to NEG-SPCA. As before, this does not affect the fact that we show tight hardness and it is an interesting open problem to remove this condition.

B.9. Robust Sparse Linear Regression

In ordinary SLR, the observations $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ are independent sample-label pairs given by $y_i = \langle \beta, X_i \rangle + \eta_i$ where $X_i \sim \mathcal{N}(0, \Sigma)$, $\eta_i \sim \mathcal{N}(0, 1)$ and β is an unknown k-sparse vector with bounded ℓ_2 norm. The task is to estimate β to within ℓ_2 norm τ. When Σ

is well-conditioned, SLR is a gapless problem with the computationally efficient LASSO attaining the information-theoretically optimal sample complexity of $n = \Theta(k \log d/\tau^2)$ (Tibshirani, 1996; Bickel et al., 2009; Raskutti et al., 2010). When Σ is not well-conditioned, SLR has a statistical-computational gap based on its restricted eigenvalue constant (Zhang et al., 2014). As with robust sparse mean estimation, the robust SLR problem $\text{RSLR}(n, k, d, \tau, \epsilon)$ is obtained when a computationally-unbounded adversary corrupts an arbitrary ϵ-fraction of the observed sample-label pairs. In this work, we consider the simplest case of $\Sigma = I_d$ where SLR is gapless but, as we discuss next, robustness seems to induce a statistical-computational gap.

Robust regression is a well-studied classical problem in statistics (Rousseeuw and Leroy, 2005). Efficient algorithms remained elusive for decades, but recent breakthroughs in sum of squares algorithms (Klivans et al., 2018; Karmalkar et al., 2019; Raghavendra and Yau, 2020), filtering approaches (Diakonikolas et al., 2019b) and robust gradient descent (Chen et al., 2017a; Prasad et al., 2018; Diakonikolas et al., 2019a) have led to the first efficient algorithms with provable guarantees. A recent line of work has also studied efficient algorithms and barriers in the high-dimensional setting of robust SLR (Chen et al., 2013; Balakrishnan et al., 2017a; Liu et al., 2018, 2019). Even in the simplest case of $\Sigma = I_d$ where the covariates X_i have independent entries, the best known polynomial time algorithms suggest robust SLR has a k-to-k^2 statistical-computational gap. As shown in Gao (2020), similar to RSME, robust SLR is only information-theoretically possible if $\tau = \Omega(\epsilon)$. In Balakrishnan et al. (2017a) and Liu et al. (2018), it is shown that polynomial-time ellipsoid-based algorithms solve robust SLR with $n = \tilde{\Theta}(k^2 \log d/\epsilon^2)$ samples when $\tau = \tilde{\Theta}(\epsilon)$. Furthermore, Liu et al. (2018) shows that an RSME oracle can be used to solve robust SLR with only a $\tilde{\Theta}(1)$ factor loss in τ and the required number of samples n. As noted in Li (2017), $n = \Omega(k \log d/\epsilon^2)$ samples suffice to solve RSME inefficiently when $\tau = \Theta(\epsilon)$. Combining these observations yields an inefficient algorithm for robust SLR with sample complexity $n = \tilde{\Theta}(k \log d/\epsilon^2)$ samples when $\tau = \tilde{\Theta}(\epsilon)$, confirming that the best known efficient algorithms suggest a k-to-k^2 statistical-computational gap. In Chen et al. (2013) and Liu et al. (2019), efficient algorithms are shown to succeed in an alternative regime where $n = \tilde{\Theta}(k \log d)$, $\epsilon = \tilde{O}(1/\sqrt{k})$ and $\tau = \tilde{O}(\epsilon\sqrt{k})$.

All of these algorithms suggest that the correct computational sample complexity for robust SLR is $n = \tilde{\Omega}(k^2\epsilon^2/\tau^4)$. In Section L.3, we deduce the following tight computational lower bound for RSLR providing evidence for this conjecture.

Theorem 9 (Lower Bounds for RSLR) *If k, d and n are polynomial in each other, $k = o(n^{1/6})$, $k = o(\sqrt{d})$ and $\epsilon < 1/2$ is such that $\epsilon = \tilde{\Omega}(n^{-1/2})$, then the k-BPC conjecture implies that there is a computational lower bound for $\text{RSLR}(n, k, d, \tau, \epsilon)$ at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

We present the reductions to MSLR and RSLR together as a single unified reduction k-PDS-TO-MSLR in Section I. As is discussed in Section L.3, MSLR and RSLR are obtained by setting $r = \epsilon^{-1} = 2$ and $\epsilon < 1/2$, respectively. The theorem above follows from a slightly modified version of this reduction, k-PDS-TO-MSLR$_R$, that removes the technical condition (T) that otherwise arises in applying k-PDS-TO-MSLR with $r = n^{\Omega(1)}$. This turns out to be more important here than in the context of RSME because, as in the reduction to MSLR, this reduction to RSLR inherits the technical condition that $k = o(n^{1/6})$ from our reduction to NEG-SPCA. This condition implicitly imposes a restriction on ϵ to satisfy that $\epsilon = \tilde{O}(n^{-1/3})$, since $\tau = \Omega(\epsilon)$ must be true for the problem to not be information-theoretically impossible. Thus our regime of interest for RSLR is a regime where the technical condition (T) is nontrivial.

As in the case of MSLR and NEG-SPCA, we emphasize that the condition $k = o(n^{1/6})$ does not affect the tightness of our lower bounds, merely restricting their regime of application. In particular, the theorem above yields a tight nontrivial statistical-computational gap in the entire parameter regime when $k = o(n^{1/6})$, $\tau = \Omega(\epsilon)$ and $\epsilon = \tilde{\Theta}(n^{-c})$ where c is any constant in the interval $[1/3, 1/2]$. We remark that the condition $k = o(n^{1/6})$ seems to be an artefact of our techniques rather than necessary.

In the context of RSLR, we view our main contribution as a set of reduction techniques relating PC_ρ to the very differently structured problem RSLR, rather than the resulting computation lower bound itself. A byproduct of our reduction is the explicit construction of an adversary modifying an ϵ-fraction of the samples in robust SLR that produces the k-to-k^2 statistical-computational gap in the theorem above. This adversary turns out to be surprisingly nontrivial on its own, but is a direct consequence of the structure of the reduction. This is discussed in detail in Sections I.2 and L.3.

B.10. Tensor Principal Component Analysis

In Tensor PCA, the observation is a single order s tensor T with dimensions $n^{\otimes s} = n \times n \times \cdots \times n$ given by $T \sim \theta v^{\otimes s} + \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$, where v has a Rademacher prior and is distributed uniformly over $\{-1, 1\}^n$ (Richard and Montanari, 2014). The task is to recover v within nontrivial ℓ_2 error $o(\sqrt{n})$ and is only information-theoretically possible if $\theta = \tilde{\omega}\left(n^{(1-s)/2}\right)$ (Richard and Montanari, 2014; Lesieur et al., 2017; Chen et al., 2018; Jagannath et al., 2018; Chen, 2019; Perry et al., 2020), in which case v can be recovered through exhaustive search. The best known polynomial-time algorithms all require the higher signal strength $\theta = \tilde{\Omega}(n^{-s/4})$, at which point v can be recovered through spectral algorithms (Richard and Montanari, 2014), the sum of squares hierarchy (Hopkins et al., 2015, 2016) and spectral algorithms based on the Kikuchi hierarchy (Wein et al., 2019). Lower bounds up to this conjectured computational barrier have been shown in the sum of squares hierarchy (Hopkins et al., 2015, 2017) and for low-degree polynomials (Kunisky et al., 2019). A number of natural "local" algorithms have also been shown to fail given much stronger levels of signal up to $\theta = \tilde{o}(n^{-1/2})$, including approximate message passing, the tensor power method, Langevin dynamics and gradient descent (Richard and Montanari, 2014; Anandkumar et al., 2014; Ben Arous et al., 2018).

We give a reduction showing that the PC_ρ conjecture implies an optimal computational lower bound at $\theta = \tilde{\Omega}(n^{-s/4})$ for tensor PCA. We show this lower bound against efficient algorithms with a low false positive probability of error in the hypothesis testing formulation of tensor PCA where $T \sim \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$ under H_0 and T is sampled from the tensor PCA distribution described above under H_1. More precisely, we prove the following theorem in Sections J and N.

Theorem 10 (Lower Bounds for TPCA) *Let n be a parameter and $s \geq 3$ be a constant, then the k-HPC^s conjecture implies a computational lower bound for* $\text{TPCA}^s(n, \theta)$ *when $\theta = \tilde{o}(n^{-s/4})$ against* $\text{poly}(n)$ *time algorithms $\mathcal{A}$ solving* $\text{TPCA}^s(n, \theta)$ *with a low false positive probability of $\mathbb{P}_{H_0}[\mathcal{A}(T) = H_1] = O(n^{-s})$.*

Lemma 102 in Section N shows that any $\text{poly}(n)$ time algorithm solving the recovery formulation of tensor PCA yields such an algorithm $\mathcal{A}$, and thus this theorem implies our desired computational lower bound. This low false positive probability of error condition on $\mathcal{A}$ arises from the fact that our reduction to TPCA is a *multi-query* average-case reduction, requiring multiple calls to a tensor PCA blackbox to solve k-HPC^s. This feature is a departure from the rest of our reductions and

the other average-case reductions to statistical problems in the literature, all of which are reductions in total variation, as will be described in Section E.2, and thus only require a single query. This feature is a requirement of our technique for completing hypergraphs that will be described further in Sections C.6 and J.

We note that most formulations of tensor PCA in the literature also assume that the noise tensor of standard Gaussians is symmetric (Richard and Montanari, 2014; Wein et al., 2019). However, given that the planted rank-1 component $v^{\otimes s}$ is symmetric as it is in our formulation, the symmetric and asymmetric noise models have a simple equivalence up to a constant factor loss in θ. Averaging the entries of the asymmetric model over all permutations of its s coordinates shows one direction of this equivalence, and the other is achieved by reversing this averaging procedure through Gaussian cloning as in Section 10 of Brennan et al. (2018). A closely related work is that of Zhang and Xia (2017), which gives a reduction from HPC^3 to the problem of detecting a planted rank-1 component in a 3-tensor of Gaussian noise. Aside being obtained through different techniques, their result differs from ours in two ways: (1) the rank-1 components they considered were sparse, rather than sampled from a Rademacher prior; and (2) their reduction necessarily produces asymmetric rank-1 components. Although the limits of tensor PCA when $s \geq 3$ with sparse and Rademacher priors are similar, they can be very different in other problems. For example, in the matrix case when $s = 2$, a sparse prior yields a problem with a statistical-computational gap while a Rademacher prior does not. We also remark that ensuring the symmetry of the planted rank-1 component is a technically difficult step and part of the motivation for our completing hypergraphs technique in Section J.

B.11. Universality for Learning Sparse Mixtures

When $\epsilon = 1/2$, our reduction to robust sparse mean estimation also implicitly shows tight computational lower bounds at $n = \tilde{o}(k^2/\tau^4)$ for learning sparse Gaussian mixtures. In this problem the task is to estimate two vectors μ_1, μ_2 up to ℓ_2 error τ, where the μ_i have bounded ℓ_2 norms and a k-sparse difference $\mu_1 - \mu_2$, given samples from an even mixture of $\mathcal{N}(\mu_1, I_d)$ and $\mathcal{N}(\mu_2, I_d)$. In general, learning in Gaussian mixture models with sparsity has been studied extensively over the past two decades (Raftery and Dean, 2006; Pan and Shen, 2007; Maugis et al., 2009; Maugis and Michel, 2011; Azizyan et al., 2013, 2015; Malsiner-Walli et al., 2016; Verzelen and Arias-Castro, 2017; Fan et al., 2018). Recent work has established finite-sample guarantees for efficient and inefficient algorithms and proven information-theoretic lower bounds for the two-component case (Azizyan et al., 2013; Verzelen and Arias-Castro, 2017; Fan et al., 2018). These works conjectured that this problem has the k-to-k^2 statistical-computational gap shown by our reduction. In Fan et al. (2018), a tight computational lower bound matching ours was established in the SQ model.

So far, despite having a variety of different hidden structures, the problems we have considered have all had either Gaussian or Bernoulli noise distributions. As we will describe in Section 4, our techniques also crucially use a number of properties of the Gaussian distribution. This naturally raises the question: do our techniques have implications beyond simple noise distributions? Our final reduction answers this affirmatively, showing that our lower bound for learning sparse Gaussian mixtures implies computational lower bounds for a wide universality class of noise distributions. This lower bound includes the optimal gap in learning sparse Gaussian mixtures and the optimal gaps in Berthet and Rigollet (2013b), Berthet and Rigollet (2013a), Wang et al. (2016b), Gao et al. (2017) and Brennan et al. (2018) for sparse PCA as special cases. This reduction requires intro-

ducing a new type of rejection kernel, that we refer to as symmetric 3-ary rejection kernels, and is described in Sections C.7 and F.3.

In Section O, we show computational lower bounds for the *generalized learning sparse mixtures* problem GLSM. In $\text{GLSM}(n, k, d, \mathcal{U})$ where $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})$, the elements of the family $\{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}$ and $\mathcal{Q}$ are distributions on a measurable space, such that the pairs $(\mathcal{P}_\nu, \mathcal{Q})$ all satisfy mild conditions permitting efficient computation outlined in Section F.3, and $\mathcal{D}$ is a mixture distribution on $\mathbb{R}$. The observations in GLSM are n independent samples $X_1, X_2, \ldots, X_n$ formed as follows:

- for each sample X_i, draw some latent variable $\nu_i \sim \mathcal{D}$ and
- sample $(X_i)_j \sim \mathcal{P}_{\nu_i}$ if $j \in S$ and $(X_i)_j \sim \mathcal{Q}$ otherwise, independently

where S is some unknown subset containing k of the d coordinates. The task is to recover S or distinguish from an H_0 in which all of the data is drawn i.i.d. from $\mathcal{Q}$. Given a collection of distribution $\mathcal{U}$, we define $\mathcal{U}$ to be in our universality class $\text{UC}(N)$ with level of signal $\tau_{\mathcal{U}}$ if it satisfies the following conditions.

Definition 13 (Universality Class and Level of Signal) *Given a parameter N, define the collection of distributions $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})$ implicitly parameterized by N to be in the universality class $\text{UC}(N)$ if*

- *the pairs $(\mathcal{P}_\nu, \mathcal{Q})$ are all computable pairs, as in Definition 24, for all $\nu \in \mathbb{R}$;*
- *$\mathcal{D}$ is a symmetric distribution about zero and $\mathbb{P}_{\nu\sim\mathcal{D}}[\nu \in [-1, 1]] = 1 - o(N^{-1})$; and*
- *there is a level of signal $\tau_{\mathcal{U}} \in \mathbb{R}$ such that for all $\nu \in [-1, 1]$ such that for any fixed constant $K > 0$, it holds that*

$$\left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x)\right| = O_N\left(\tau_{\mathcal{U}}\right) \quad \text{and} \quad \left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) - 2\right| = O_N\left(\tau_{\mathcal{U}}^2\right)$$

with probability at least $1 - O\left(N^{-K}\right)$ over each of $\mathcal{P}_\nu, \mathcal{P}_{-\nu}$ and $\mathcal{Q}$.

Our main result establishes a computational lower bound for GLSM instances with $\mathcal{U} \in \text{UC}(n)$ in terms of the level of signal $\tau_{\mathcal{U}}$. As mentioned above, this theorem implies optimal lower bounds for learning sparse mixtures of Gaussians, sparse PCA and many more natural problem formulations described in Section O.2.

Theorem 11 (Computational Lower Bounds for GLSM) *Let n, k and d be polynomial in each other and such that $k = o(\sqrt{d})$. Suppose that the collections of distributions $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})$ is in $\text{UC}(n)$. Then the k-BPC conjecture implies a computational lower bound for $\text{GLSM}(n, k, d, \mathcal{U})$ at all sample complexities $n = \tilde{o}\left(\tau_{\mathcal{U}}^{-4}\right)$.*

Appendix C. Detailed Technical Overview

We now outline our main technical contributions and the central ideas behind our reductions, expanding significantly on the outline in Section 4. These techniques will be formally introduced in Part II and applied in our problem-specific reductions to deduce our main theorems stated in the previous section in Part III.

C.1. Rejection Kernels

Rejection kernels are a reduction primitive introduced in Brennan et al. (2018, 2019a) for *algorithmic changes of measure*. Related reduction primitives for changes of measure to Gaussians and binomial random variables appeared earlier in Ma and Wu (2015) and Hajek et al. (2015). Given two input Bernoulli probabilities $0 < q < p \leq 1$, a rejection kernel simultaneously maps $\text{Bern}(p)$ and $\text{Bern}(q)$ approximately in total variation to samples from two arbitrary distributions $\mathcal{P}$ and $\mathcal{Q}$. Note that in this setup, the rejection kernel primitive is oblivious to whether the true distribution of its input is $\text{Bern}(p)$ or $\text{Bern}(q)$. The main idea behind rejection kernels is that, under suitable conditions on $\mathcal{P}$ and $\mathcal{Q}$, this can be achieved through a rejection sampling scheme that samples $x \sim \mathcal{Q}$ and rejects with a probability that depends on x and on whether the input was 0 or 1. Rejection kernels are discussed in more depth in Section F. In this work, we will need the following two instantiations of the framework developed in Brennan et al. (2018, 2019a):

- *Gaussian Rejection Kernels:* Rejection kernels mapping $\text{Bern}(p)$ and $\text{Bern}(q)$ to within $O(R_{\text{RK}})$ total variation of $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(0, 1)$ where $\mu = \Theta\left(1/\sqrt{\log R_{\text{rk}}^{-1}}\right)$ and p, q are fixed constants.
- *Bernoulli Cloning:* A rejection kernel mapping $\text{Bern}(p)$ and $\text{Bern}(q)$ exactly to $\text{Bern}(P)^{\otimes t}$ and $\text{Bern}(Q)^{\otimes t}$ where

$$\frac{1-p}{1-q} \leq \left(\frac{1-P}{1-Q}\right)^t \quad \text{and} \quad \left(\frac{P}{Q}\right)^t \leq \frac{p}{q}$$

By performing computational changes of measure, these primitives are crucial in mapping to desired distributional aesthetics. However, they also play an important role in transforming hidden structure. Gaussian rejection kernels grant access to an arsenal of measure-preserving transformations of high-dimensional Gaussian vectors for mapping between different hidden structures while preserving independence in the noise distribution. Bernoulli cloning is crucial in removing the symmetry in adjacency matrices of PC instances and adjacency tensors of HPC instances, as in the TO-SUBMATRIX procedure in Brennan et al. (2019a). We introduce a k-partite variant of this procedure that maps the adjacency matrix of k-PDS to a matrix of independent Bernoulli random variables while respecting the constraint that there is one planted entry per block of the k-partition. This procedure is discussed in more detail in Section C.6 and will serve as a crucial preprocessing step for dense Bernoulli rotations, which involves taking linear combinations of functions of entries of this matrix that crucially must be independent.

C.2. Dense Bernoulli Rotations

This technique is introduced in Section G and is one of our main primitives for *transforming hidden structure* that will be applied repeatedly throughout our reductions. Let $\text{PB}(n, i, p, q)$ denote the planted bit distribution over $V \in \{0,1\}^n$ with independent entries satisfying that $V_j \sim \text{Bern}(q)$ unless $j = i$, in which case $V_i \sim \text{Bern}(p)$. Given an input vector $V \in \{0,1\}^n$, the goal of dense Bernoulli rotations is to output a vector $V' \in \mathbb{R}^m$ such that, for each $i \in [n]$, V' is close in total variation to $\mathcal{N}(c \cdot A_i, I_m)$ if $V \sim \text{PB}(n, i, p, q)$. Here, $A_1, A_2, \ldots, A_n \in \mathbb{R}^m$ are a given sequence of target mean vectors, p and q are fixed constants and c is a scaling factor with $c = \tilde{\Theta}(1)$. The

reduction must satisfy these approximate Markov transition conditions oblivious to the planted bit i and also preserve independent noise, by mapping $\text{Bern}(q)^{\otimes n}$ to $\mathcal{N}(0, I_m)$ approximately in total variation.

Let $A \in \mathbb{R}^{m \times n}$ denote the matrix with columns $A_1, A_2, \ldots, A_n$. If the rows of A are orthogonal unit vectors, then the goal outlined above can be achieved using the isotropy of the distribution $\mathcal{N}(0, I_n)$. More precisely, consider the reduction that form $V_1 \in \mathbb{R}^n$ by applying Gaussian rejection kernels entrywise to V and then outputs AV_1. If $V \sim \text{PB}(n, i, p, q)$, then the rejection kernels ensure that V_1 is close in total variation to $\mathcal{N}(\mu \cdot \mathbf{1}_i, I_n)$ and thus $V' = AV_1$ is close to $\mathcal{N}(\mu \cdot A_i, I_m)$. However, if the rows of A are not orthogonal, then the entries of the output are potentially very dependent and have covariance matrix $AA^\top$ instead of I_m. This can be remedied by adding a *noise-correction term* to the output: generate $U \sim \mathcal{N}(0, I_m)$ and instead output

$$V' = \lambda^{-1} \cdot AV_1 + \left(I_m - \lambda^{-2} \cdot AA^\top\right)^{1/2} \cdot U$$

where λ is an upper bound on the largest singular value of A and $\left(I_m - \lambda^{-2} AA^\top\right)^{1/2}$ is the positive semidefinite square root of $I_m - \lambda^{-2} \cdot AA^\top$. If $V \sim \text{PB}(n, i, p, q)$, it now follows that V' is close in total variation to $\mathcal{N}(\mu\lambda^{-1} \cdot A_i, I_m)$ where μ can be taken to be $\mu = \Theta(1/\sqrt{\log n})$. This reduction also preserves independent noise, mapping $\text{Bern}(q)^{\otimes n}$ approximately to $\mathcal{N}(0, I_m)$.

Dense Bernoulli rotations thus begin with a random vector of independent entries and one unknown elevated bit and produce a vector with independent entries and an unknown elevated *pattern* from among an arbitrary prescribed set $A_1, A_2, \ldots, A_n$. Furthermore, the dependence of the signal strength $\mu\lambda^{-1}$ in the output instance V' on these $A_1, A_2, \ldots, A_n$ is entirely through the singular values of A. This yields a general structure-transforming primitive that will be used throughout our reductions. Each such use will consist of many local applications of dense Bernoulli rotations that will be stitched together to produce a target distribution. These local applications will take three forms:

- *To Rows Restricted to Column Parts:* The adjacency matrix of k-BPC consists of $k_n k_m$ blocks each consisting of the edge indicators in $E_i \times F_j$ for each pair of the parts E_i, F_j from the given partitions of $[n]$ and $[m]$. In our reductions to robust sparse mean estimation, mixtures of SLRs, robust SLR and universality for learning sparse mixtures, we apply dense Bernoulli rotations separately to each row in each of these blocks.

- *To Vectorized Adjacency Matrix Blocks:* In our reductions to dense stochastic block models, testing hidden partition models and semirandom single community detection, we first pre-process the adjacency matrix of k-PC with TO-k-PARTITE-SUBMATRIX. We then apply dense Bernoulli rotations to $\mathbb{R}^{h^2}$ vectorizations of each $h \times h$ block in this matrix corresponding a pair of parts in the given partition i.e. of the form $E_i \times E_j$.

- *To Vectorized Adjacency Tensor Blocks:* In our reduction to tensor PCA with order s, after completing the adjacency tensor of the input k-HPC instance, we apply dense Bernoulli rotations to $\mathbb{R}^{h^s}$ vectorizations of each $h \times h \times \cdots \times h$ block corresponding to an s-tuple of parts.

We remark that while dense Bernoulli rotations heavily rely on distributional properties of isotropic Gaussian vectors, their implications extend far beyond statistical problems with Gaussian noise.

Entrywise thresholding produces planted graph problems and we will show that multiple thresholds followed by applying 3-ary symmetric rejection kernels maps to a large universality class of noise distributions. These applications of dense Bernoulli rotations generally reduce the problem of transforming hidden structure to a constrained combinatorial construction problem – the task of designing a set of mean output vectors $A_1, A_2, \ldots, A_n$ that have nearly orthogonal rows and match the combinatorial structure in the target statistical problem.

C.3. Design Matrices and Tensors

Design Matrices. To construct these vectors $A_1, A_2, \ldots, A_n$ for our applications of dense Bernoulli rotations, we introduce several families of matrices based on the incidence geometry of finite fields. In our reduction to robust sparse mean estimation, we will show that the adversary that corrupts an ϵ-fraction of the samples by resampling them from $\mathcal{N}(-c \cdot \mu, I_d)$ produces the desired k-to-k^2 statistical-computational gap. This same adversarial construction was used in Diakonikolas et al. (2017). Here, $\mu \in \mathbb{R}^d$ denotes the k-sparse mean of interest. As will be further discussed at the beginning of Section G, on applying dense Bernoulli rotations to rows restricted to parts of the partition of column partition, our desiderata for the mean vectors $A_1, A_2, \ldots, A_n$ reduce to the following:

- A contains two distinct values $\{x, y\}$, and an ϵ'-fraction of each column is y where $\epsilon \geq \epsilon' = \Theta(\epsilon)$;
- the rows of A are unit vectors and nearly orthogonal with $\lambda = O(1)$; and
- A is nearly an isometry as a linear transformation from $\mathbb{R}^n \to \mathbb{R}^m$.

The first criterion above is enough to ensure the correct distributional aesthetics and hidden structure in the output of our reduction. The second and third criteria turn out to be necessary and sufficient for the reduction to show tight computational lower bounds up to the conjectured barrier of $n = \tilde{o}(k^2\epsilon^2/\tau^4)$. We remark that the third criterion also is equivalent to $m = \tilde{\Theta}(n)$ given the second. Thus our task is to design nearly square, nearly orthogonal matrices containing two distinct entries with an ϵ'-fraction of one present in each column. Note that if $\epsilon = 1/2$, this is exactly achieved by Hadamard matrices. For $\epsilon < 1/2$, our desiderata are nearly met by the following natural generalization of Hadamard matrices that we introduce. Note that the rows of a Hadamard matrix can be generated as a reweighted incidence matrix between the hyperplanes and points of $\mathbb{F}_2^t$. Let r be a prime number with $\epsilon^{-1} \leq r = O(\epsilon^{-1})$ and consider the $\ell \times r^t$ matrix A where $\ell = \frac{r^t - 1}{r-1}$ with entries given by

$$A_{ij} = \frac{1}{\sqrt{r^t(r-1)}} \cdot \begin{cases} 1 & \text{if } P_j \notin V_i \\ 1 - r & \text{if } P_j \in V_i \end{cases}$$

where $V_1, V_2, \ldots, V_\ell$ is an enumeration of the $(t-1)$-dimensional subspaces of $\mathbb{F}_r^t$ and $P_1, P_2, \ldots, P_{r^t}$ is an enumeration of the points in $\mathbb{F}_r^t$. This construction nearly meets our three criteria, with one minor issue that the column corresponding to $0 \in \mathbb{F}_r^t$ only contains one entry. A more serious issue is that $\ell = \Theta(r^{t-1})$ and A is far from an isometry if $r \gg 1$, which leads to a suboptimal computational lower bound for RSME.

These issues are both remedied by adding in additional rows for all affine shifts of the hyperplanes $V_1, V_2, \ldots, V_\ell$. The resulting matrix has dimensions $r\ell \times r^t$ and, although its rows are no

longer orthogonal, its largest singular value is $\sqrt{1+(r-1)^{-1}}$. The resulting matrix $K_{r,t}$ is used in our applications of dense Bernoulli rotations to reduce to robust sparse mean estimation, mixtures of SLRs, robust SLR and to show universality for learning sparse mixtures. Note that for any two rows r_i and r_j of $K_{r,t}$, the outer product $r_i r_j^\top$ is a zero-centered mean adjacency matrix of an imbalanced 2-block stochastic block model. This observation suggests that the Kronecker product $K_{r,t} \otimes K_{r,t}$ can be used in dense Bernoulli rotations to map to these SBMs. Surprisingly, this overall reduction yields tight computational lower bounds up to the Kesten-Stigum threshold for dense SBMs, and using the matrix $(K_{3,t} \otimes I_s) \otimes (K_{3,t} \otimes I_s)$ yields tight computational lower bounds for semirandom single community detection. We remark that, in this case, it is again crucial that $K_{r,t}$ is approximately square – if the matrix A defined above were used in place of $K_{r,t}$, our reduction would show a lower bound suboptimal to the Kesten-Stigum threshold by a factor of r. Our reduction to order s tensor PCA applies dense Bernoulli rotations to vectorizations of each tensor block with the sth order Kronecker product $K_{2,t} \otimes K_{2,t} \otimes \cdots \otimes K_{2,t}$. We remark that these instances of $K_{2,t}$ in this Kronecker product could be replaced by Hadamard matrices in dimension 2^t.

In Section G.4, we introduce a natural alternative to $K_{r,t}$ – a random matrix $R_{n,\epsilon}$ that *approximately* satisfies the three desiderata above. In our reductions to RSME and RSLR, this random matrix has the advantage of eliminating the number-theoretic condition (T) arising from applying dense Bernoulli rotations with $K_{r,t}$, which has nontrivial restrictions in the very small ϵ regime when $\epsilon = n^{-\Omega(1)}$. However, the approximate properties of $R_{n,\epsilon}$ are insufficient to map exactly to our formulations of ISBM, SEMI-CR, GHPM and BHPM, where the sizes of the hidden communities are known. A more detailed comparison of $K_{r,t}$ and $R_{n,\epsilon}$ can be found in Section G.4. The random matrix $R_{n,\epsilon}$ is closely related to the adjacency matrices of sparse random graphs, and establishing $\lambda = O(1)$ requires results on their spectral concentration from the literature. For a consistent and self-contained exposition, we present our reductions with $K_{r,t}$, which has a comparatively simple analysis, and only outline extensions of our reductions using $R_{n,\epsilon}$.

Design Tensors. Our final reduction using dense Bernoulli rotations is to testing hidden partition models. This reduction requires a more involved construction for A that we only sketch here and defer a detailed discussion to Section G.3. Again applying dense Bernoulli rotations to vectorizations of each block of the input k-PC instance, our goal is to construct a tensor $T_{r,t}$ such that each slice has the same block structure as an r-block SBM and the slices are approximately orthogonal under the matrix inner product. A natural construction is as follows: index each slice by a pair of hyperplanes (V_i, V_j), label the rows and columns of each slice by $\mathbb{F}_r^t$ and plant r communities on the entries with indices in $(V_i + au_i) \times (V_j + au_j)$ for each $a \in \mathbb{F}_r$. Here u_i and u_j are arbitrary vectors not in V_i and V_j, respectively, and thus $V_i + au_i$ ranges over all affine shifts of V_i for $a \in \mathbb{F}_r$. An appropriate choice of weights x and y on and off of these communities yields slices that are exactly orthogonal.

However, this construction suffers from the same issue as the construction of A above – there are $O(r^{2t-2})$ slices each of which has r^{2t} entries, making the matrix formed by vectorizing the slices of this tensor far from square. This can be remedied by creating additional slices further indexed by a nonconstant linear function $L : \mathbb{F}_r \to \mathbb{F}_r$ such that communities are now planted on $(V_i + au_i) \times (V_j + L(a) \cdot u_j)$ for each $a \in \mathbb{F}_r$. There are $r(r-1)$ such linear functions L, making the vectorization of this tensor nearly square. Furthermore, it is shown in Section G.3 that this matrix has largest singular value $\sqrt{1+(r-1)^{-1}}$. We remark that this property is quite brittle, as substituting other families of bijections for L can cause this largest singular value to increase

dramatically. Taking the Kronecker product of each slice of this tensor $T_{r,t}$ with I_s now yields the family of matrices used in our reduction to testing hidden partition models.

We remark that in all of these reductions with both design matrices and design tensors, dense Bernoulli rotations are applied locally within the blocks induced by the partition accompanying the PC_ρ instance. In all cases, our constructions ensure that the fact that the planted bits within these blocks take the form of a submatrix is sufficient to stitch together the outputs of these local applications of dense Bernoulli rotations into a single instance with the desired hidden structure. While we did not discuss this constraint in choosing the design matrices A for each of our reductions, it will be a key consideration in the proofs throughout this work. Surprisingly, the linear functions L in the construction of $T_{r,t}$ directly lead to a community alignment property proven in Section G.3 that allow slices of this tensor to be consistently stitched together. Furthermore, we note that unlike $K_{r,t}$, the tensor $T_{r,t}$ does not seem to have a random matrix analogue that is tractable to bound in spectral norm.

Parameter Correspondence with Dense Bernoulli Rotations. In several of our reductions using dense Bernoulli rotations, a simple heuristic predicts our computational lower bound in the target problem. Let X be a data tensor, normalized and centered so that each entry has mean zero and variance 1, and then consider the ℓ_2 norm of the expected tensor $\mathbb{E}[X]$. Our applications of rejection kernels typically preserve this ℓ_2 norm up to $\text{polylog}(n)$ factors. Since our design matrices are approximate isometries, most of our applications of dense Bernoulli rotations also approximately preserve this ℓ_2 norm. Thus comparing the ℓ_2 norms of the input PC_ρ instance and output instance in our reductions yields a heuristic for predicting the resulting computational lower bound. For example, our adversary in RSME produces a matrix $\mathbb{E}[X] \in \mathbb{R}^{d \times n}$ consisting of columns of the form $\tau \cdot k^{-1/2} \cdot \mathbf{1}_S$ and $\epsilon^{-1}(1-\epsilon)\tau \cdot k^{-1/2} \cdot \mathbf{1}_S$, up to constant factors where S is the hidden support of μ. The ℓ_2 norm of this matrix is $\Theta(\tau\sqrt{n/\epsilon})$. The ℓ_2 norm of the matrix $\mathbb{E}[X]$ corresponding to the starting k-BPC instance can be verified to be just below $o(k^{1/2}n^{1/4})$, when the k-BPC instance is nearly at its computational barrier. Equating these two ℓ_2 norms yields the relation $n = \Theta(k^2\epsilon^2/\tau^4)$, which is exactly our computational barrier for RSME. Similar heuristic derivations of our computational barriers are produced for ISBM, GHPM, BHPM, SEMI-CR and TPCA at the beginnings of Sections M and N. We remark that for some of our problems with central steps other than dense Bernoulli rotations, such as MSLR, RSLR and GLSM, this heuristic does not apply.

C.4. Decomposing Linear Regression and Label Generation

Our reductions to mixtures of SLRs and robust SLR in Section I are motivated by the following simple initial observation. Suppose (X, y) is a single sample from unsigned SLR with $y = \gamma R \cdot \langle v, X\rangle + \mathcal{N}(0,1)$ where $R \in \{-1, 1\}$ is a Rademacher random variable, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $X \sim \mathcal{N}(0, I_d)$ and $\gamma \in (0, 1)$. A standard conditioning property of Gaussian vectors yields that the conditional distribution of X given R and y is another jointly Gaussian vector, as shown below. Our observation is that this conditional distribution can be decomposed into a sum of our adversarial construction for robust sparse mean estimation and an independent instance of negative

sparse PCA. More formally, we have that

$$X|R, y \sim \mathcal{N}\left(\frac{R\gamma \cdot y}{1+\gamma^2} \cdot v,\ I_d - \frac{\gamma^2}{1+\gamma^2} \cdot vv^\top\right)$$
$$\sim \underbrace{\frac{1}{\sqrt{2}} \cdot \mathcal{N}\left(R\tau \cdot v,\ I_d\right)}_{\text{Our RSME adversary with } \epsilon = 1/2} + \underbrace{\frac{1}{\sqrt{2}} \cdot \mathcal{N}\left(0,\ I_d - \theta vv^\top\right)}_{\text{Negative Sparse PCA}}$$

where $\tau = \tau(y) = \frac{\gamma\sqrt{2}}{1+\gamma^2} \cdot y$ and $\theta = \frac{2\gamma^2}{1+\gamma^2}$. Note that the marginal distribution of y is $\mathcal{N}(0, 1+\gamma^2)$ and thus it typically holds that $|y| = \Theta(1)$. When this unsigned SLR instance is at its computational barrier of $n = \tilde{\Theta}(k^2/\gamma^4)$ and $|y| = \Theta(1)$, then $n = \tilde{\Theta}(k^2/\tau^4)$ and $\theta = \tilde{\Theta}(\sqrt{k^2/n})$. Therefore surprisingly, both of the RSME and NEG-SPCA in the decomposition above are also at their computational barriers.

Now consider task of instead reducing from k-BPC to the problem of estimating v from n independent samples from the conditional distribution $\mathcal{L}(X\,|\,|y| = 1)$. In light of the observations above, it suffices to first use Bernoulli cloning to produce two independent copies of k-BPC, reduce these two copies as outlined below and then take the sum of the two outputs of these reductions.

- *Producing Our* RSME *Adversary*: One of the two copies of k-BPC should be mapped to a tight instance of our adversarial construction for RSME with $\epsilon = 1/2$ through local applications of dense Bernoulli rotations with design matrix $K_{r,t}$ or $R_{n,\epsilon}$, as described previously.

- *Producing* NEG-SPCA: The other copy should be mapped to a tight instance of negative sparse PCA. This requires producing negatively correlated data from positively correlated data, and will need new techniques that we discuss next.

We remark that while these two output instances must be independent, it is important that they share the same latent vector v. Bernoulli cloning ensures that the two independent copies of k-PC have the same clique vertices and thus the output instances have this desired property.

This reduction can be extended to reduce to the true joint distribution of (X, y) as follows. Consider replacing each sample X_1 of the output RSME instance by

$$X_2 = cy \cdot X_1 + \sqrt{1 - c^2y^2} \cdot \mathcal{N}(0, I_d)$$

where c is some scaling factor and y is independently sampled from $\mathcal{N}(0, 1+\gamma^2)$, truncated to lie in the interval $[-T, T]$ where $cT \le 1$. Observe that if $X_1 \sim \mathcal{N}(R\tau \cdot v, I_d)$, then $X_2 \sim \mathcal{N}(cR\tau y \cdot v, I_d)$ conditioned on y. In Section I.2, we show that a suitable choice of c, T and tweaking τ in the reduction above tightly maps to the desired distribution of mixtures of SLRs. Analogous observations and performing the RSME sub-reduction with $\epsilon < 1/2$ can be used to show tight computational lower bounds for robust SLR. We remark that this produces a more complicated adversarial construction for robust SLR that may be of independent interest. The details of this adversary can be found in Section I.2.

C.5. Producing Negative Correlations and Inverse Wishart Matrices

To complete our reductions to mixtures of SLRs and robust SLR, it suffices to give a tight reduction from k-BPC to NEG-SPCA. Although NEG-SPCA and ordinary SPCA share the same conjectured

computational barrier at $\theta = \Theta(\sqrt{k^2/n})$ and can be solved by similar efficient algorithms above this barrier, as stochastic models, the two are very different. As discussed in Section B.7, ordinary SPCA admits a signal plus noise representation while NEG-SPCA does not. This representation was crucially used in prior reductions showing optimal computational lower bounds for SPCA in Berthet and Rigollet (2013b), Berthet and Rigollet (2013a), Wang et al. (2016b), Gao et al. (2017), Brennan et al. (2018) and Brennan and Bresler (2019). Furthermore, the planted entries in a NEG-SPCA sample are *negatively correlated*. In contrast, the edge indicators of PC_ρ are positively correlated and all prior reductions from PC have only produced hidden structure that is also positively correlated.

We first simplify the task of reducing to NEG-SPCA with an observation used in the reduction to SPCA in Brennan and Bresler (2019). Suppose that $n \geq m + 1$ and let m be such that m/k^2 tends slowly to infinity. If X is an $m \times n$ matrix with columns $X_1, X_2, \ldots, X_n \sim_{\text{i.i.d.}} \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{m \times m}$ is positive semidefinite, then the conditional distribution of X given its rescaled empirical covariance matrix $\hat{\Sigma} = \sum_{i=1}^n X_i X_i^\top$ is $\hat{\Sigma}^{1/2} R$ where R is an independent $m \times n$ matrix sampled from Haar measure on the Stiefel manifold. This implies that it suffices to reduce to $\hat{\Sigma}$ in the case where $\Sigma = I_d - \theta vv^\top$ in order to map to NEG-SPCA, as X can be generated from $\hat{\Sigma}$ by randomly sampling this Haar measure. This measure can then be sampled efficiently by applying Gram-Schmidt to the rows of an $m \times n$ matrix of independent standard Gaussians.

Let $\mathcal{W}_m(n, \Sigma)$ be the law of $\hat{\Sigma}$, or in other words the Wishart distribution with covariance matrix Σ, and let $\mathcal{W}_m^{-1}(n, \Sigma)$ denote the distribution of its inverse. The matrices $\mathcal{W}_m(n, \Sigma)$ and $\mathcal{W}_m^{-1}(n, \beta \cdot \Sigma^{-1})$ where $\beta^{-1} = n(n-m-1)$ have a number of common properties including close low-order moments. Furthermore, if $\Sigma = I_d - \theta vv^\top$ then $\Sigma^{-1} = I_d + \theta' vv^\top$ where $\theta' = \frac{\theta}{1-\theta}$, which implies that $\mathcal{W}_m^{-1}(n, \beta \cdot \Sigma^{-1})$ is a rescaling of the inverse of the empirical covariance matrix of a set of samples from ordinary SPCA. This motivates our main reduction to NEG-SPCA in Section H.1, which roughly proceeds in the following two steps.

1. Begin with a small instance of BPC with $m = \omega(k^2)$ vertices on the left and n on the right. Apply either the reduction of Brennan et al. (2018) or Brennan and Bresler (2019) to reduce to an ordinary SPCA instance $(X_1, X_2, \ldots, X_n)$ in dimension m with n samples and signal strength θ'.

2. Form the rescaled empirical covariance matrix $\hat{\Sigma} = \sum_{i=1}^n X_i X_i^\top$ and

$$Y = \sqrt{n(n-m-1)} \cdot \hat{\Sigma}^{-1/2} R$$

   Output the columns of Y after padding them to be d-dimensional with i.i.d. $\mathcal{N}(0,1)$ random variables.

The key detail in this reduction is that $\hat{\Sigma}^{1/2}$ in process of regenerating X from $\hat{\Sigma}$ described above has been replaced by the positive semidefinite square root $\hat{\Sigma}^{-1/2}$ of a rescaling of the empirical precision matrix. As we will show in Section H.1, establishing total variation guarantees for this reduction amounts to answering the following nonasymptotic question from random matrix theory that may be of independent interest: when do $\mathcal{W}_m(n, \Sigma)$ and $\mathcal{W}_m^{-1}(n, \beta \cdot \Sigma^{-1})$ converge in total variation for all positive semidefinite matrices Σ? A simple reduction shows that the general case is equivalent to the isotropic case when $\Sigma = I_m$. In Section H.2, we answer this question, showing that these two matrices converge in KL divergence if and only if $n \gg m^3$. Our result is of the same flavor as a number of recent results in random matrix theory showing convergence in total variation between Wishart and GOE matrices (Jiang and Li, 2015; Bubeck et al., 2016; Bubeck and Ganguly,

2016; Rácz and Richey, 2019). This condition amounts to constraining our reduction to the low-sparsity regime $k \ll n^{1/6}$. As discussed in Section B.7, this condition does not affect the tightness of our lower bounds and seems to be an artefact of our techniques that possibly can be removed.

C.6. Completing Tensors from Hypergraphs and Tensor PCA

As alluded to in the above discussion of rejection kernels, it is important that the entries in the vectors to which we apply dense Bernoulli rotations are independent and that none of these entries is missing. In the context of reductions beginning with k-PC, k-HPC, PC and HPC, establishing this entails pre-processing steps to remove the symmetry of the input adjacency matrix and add in missing entries. As discussed in Section 1.1 of Brennan et al. (2019a), these missing entries in the matrix case have led to technical complications in the prior reductions in Hajek et al. (2015), Brennan et al. (2018, 2019a) and Brennan and Bresler (2019). In reductions to tensor PCA, completing these pre-processing steps in the tensor case seems unavoidable in order to produce the canonical formulation of tensor PCA with a symmetric rank-1 spike $v^{\otimes s}$ as discussed in Section B.10.

In order to motivate our discussion of the tensor case, we first consider the matrix case. Asymmetrizing the adjacency matrix of an input PC instance can be achieved through a simple application of Bernoulli cloning, but adding in the missing diagonal entries is more subtle. Note that the desired diagonal entries contain nontrivial information about the vertices in the planted clique – they are constrained to be 1 along the vertices of the clique and independent Bern$(1/2)$ random variables elsewhere. This is roughly the information gained on revealing a single vertex from the planted clique. In the matrix case, the following trick effectively produces an instance of PC with the diagonal entries present. Add in 1's along the entire diagonal and randomly embed the resulting matrix as a principal minor in a larger matrix with off-diagonal entries sampled from Bern$(1/2)$ and on-diagonal entries sampled so that the total number of 1's on the diagonal has the correct binomial distribution. This trick appeared in the TO-SUBMATRIX procedure in Brennan et al. (2019a) for general PDS instances, and is adapted in this work for k-PDS as the reduction TO-k-PARTITE-SUBMATRIX in Section F. This reduction is an important pre-processing step in mapping to dense stochastic block models, testing hidden partition models and semirandom planted dense subgraph.

The tensor case is not as simple as the matrix case. While asymmetrizing can be handled similarly with Bernoulli cloning, the missing entries of the adjacency tensor of HPC are now more numerous and correspond to any entry with two equal indices. Unlike in the matrix case, the information content in these entries alone is enough to solve HPC. For example, in 3-uniform HPC, the missing set of entries (i, i, j) should have the same distribution as the completed adjacency matrix of an entire instance of planted clique with the same hidden clique vertices. Thus a reduction that randomly generates these missing entries as in the matrix case is no longer possible without knowing the solution to the input HPC instance. However, if an oracle were to have revealed a single vertex of the hidden clique, we would be able to use the hyperedges containing this vertex to complete the missing entries of the adjacency tensor. In general, given an HPC instance of arbitrary order s, a more involved cloning and embedding procedure detailed in Section J completes the missing entries of the adjacency tensor given oracle access to $s - 1$ vertices of the hidden clique. Our reduction to tensor PCA in Sections J and N iterates over all $(s - 1)$-tuples of vertices in the input HPC instance, uses this procedure to complete the missing entries of the adjacency tensor, applies dense Bernoulli rotations as described previously and then feeds the output instance to a blackbox solving tensor PCA. The reduction only succeeds in mapping to the correct distribution on tensor

PCA in iterations that successfully guess $s-1$ vertices of the planted clique. However, we show that this is sufficient to deduce tight computational lower bounds for tensor PCA. We remark that this reduction is the first reduction in total variation from PC_ρ that seems to require multiple calls to a blackbox solving the target problem.

C.7. Symmetric 3-ary Rejection Kernels and Universality

So far, all of our reductions have been to problems with Gaussian or Bernoulli data and our techniques have often relied heavily on the properties of jointly Gaussian vectors. Our last reduction technique shows that the consequences of these reductions extend far beyond Gaussian and Bernoulli problems. We introduce a new rejection kernel in Section F.3 and show in Section O that, when applied entrywise to the output of our reduction to RSME when $\epsilon = 1/2$, this rejection kernel yields a universal computational lower bound for a general variant of learning sparse mixtures with nearly arbitrary marginals.

Because sparse mixture models necessarily involve at least three distinct marginal distributions, a deficit in degrees of freedom implies that the existing framework for rejection kernels with binary entries cannot yield nontrivial hardness. We resolve this issue by considering rejection kernels with a slightly larger input space, and introduce a general framework for 3-ary rejection kernels with entries in $\{-1, 0, 1\}$ in Section F.3. We show in Section O that first mapping each entry of our RSME instance with $\epsilon = 1/2$ into $\{-1, 0, 1\}$ by thresholding at intervals of the form $(-\infty, -T], (-T, T)$ and $[T, \infty)$ with $T = \Theta(1)$ and then applying 3-ary rejection kernels entrywise is a nearly lossless reduction. In particular, it yields new computational lower bounds for a wide universality class that tightly recover optimal computational lower bounds for sparse PCA, learning mixtures of exponentially distributed data, the original RSME instance with $\epsilon = 1/2$ and many other sparse mixture formulations. The implications of this reduction are discussed in detail in Section O.2.

C.8. Encoding Cliques as Structural Priors

As discussed in Section A.2, reductions from PC_ρ showing tight computational lower bounds cannot generate a non-negligible part of the hidden structure in the target problem themselves, but instead must encode the hidden clique of the input instance into this structure. In this section, we outline how our reductions implicitly encode hidden cliques. Note that the hidden subset of vertices corresponding to a clique in PC_ρ has $\Theta(k \log n)$ bits of entropy while the distribution over the hidden structure in the target problems that we consider can have much higher entropy. For example, the Rademacher prior on the planted vector v in Tensor PCA has n bits of entropy and the distribution over hidden partitions in testing partition models has entropy $\Theta(r^2 K^2 \log n \log r)$.

Although our reductions inject randomness to produce the desired noise distributions of target problems, the induced maps encoding the clique in PC_ρ as a new hidden structure typically do not inject randomness. Consequently, our reductions generally show hardness for priors over the hidden structure in our target problems with entropy $\Theta(k \log n)$. This then implies a lower bound for our target problems, because the canonical uniform priors with which they are defined are the *hardest priors*. For example, every instance of PC_ρ reduces to uniform prior over cliques as in PC by randomly relabelling nodes. Similarly, a tensor PCA instance with a fixed planted vector v reduces to the formulation in which v is uniformly distributed on $\{-1, 1\}^n$ by taking the entrywise product of the tensor PCA instance with $u^{\otimes s}$ where u is chosen u.a.r. from $\{-1, 1\}^n$. Thus our reductions actually show slightly stronger computational lower bounds than those stated in our main theorems –

they show lower bounds for our target problems with *nonuniform* priors on their hidden structures. These nonuniform priors arise from the encodings of planted cliques into target hidden structure implicitly in our reductions, several of which we summarize below. Our reductions often involve aesthetic pre-processing and post-processing steps to reduce to canonical uniform priors and often subsample the output instance. To simplify our discussion, we omit these steps in describing the clique encodings induced by our reductions.

- **Robust Sparse Mean Estimation and SLR:** Let S_L and S_R be the sets of left and right clique vertices of the input k-BPC instance and let $[N] = E_1 \cup E_2 \cup \cdots \cup E_{k_N}$ be the given partition of the right vertices. The support of the k-sparse vector in our output RSME and RSLR instances is simply S_L. Let r be a prime and let $E'_1 \cup E'_2 \cup \cdots \cup E'_{k_N}$ be a partition of the output n samples into parts of size $r\ell$ where $\ell = \frac{r^t-1}{r-1}$. Label each of element of E'_i with a affine shift of a hyperplane in $\mathbb{F}_r^t$ and each element of E_i with a point of $\mathbb{F}_r^t$. For each i, our adversary corrupts each sample in E'_i corresponding to an affine shift of a hyperplane containing the point corresponding to the unique element in $S_R \cap E_i$.

- **Dense Stochastic Block Models:** Let S be the set of clique vertices of the input k-PC instance and let E be the given partition of the its vertices $[N]$. Let E' be a partition of the output n vertices again into parts of size $r\ell$. Label elements in each part as above. Our output ISBM instance has its smaller community supported on the union of the vertices across all E'_i corresponding to affine shifts containing the points in $\mathbb{F}_r^t$ corresponding to the vertices S.

- **Mixtures of SLRs and Generalized Learning Sparse Mixtures:** Let S_L, S_R, k, k_N, N, n and E be as above. The support of the k-sparse vector in our output MSLR and GLSM instances is again simply S_L. Let $H_1, H_2, \ldots, H_{2^t-1} \in \{-1, 1\}^{2^t}$ be the zero-sum rows of a Hadamard matrix and let E' be a partition of the output n samples into k_N blocks of size 2^t. The output instance sets the jth sample in E'_i to be from the first part of the mixture if and only if the jth entry of H_s is 1 where s is the unique element in $S_R \cap E_i$. In other words, the mixture pattern along E'_i is given by the $(S_R \cap E_i)$th row of a Hadamard matrix.

- **Tensor PCA:** Let S be the set of clique vertices of the input k-HPC instance and let E and N be as above. Similarly to MSLR and GLSM, the planted vector v of our output TPCA instance is the concatenation of the $(S \cap E_i)$th rows of a Hadamard matrix.

Our reduction to testing hidden partition models induces a more intricate encoding of cliques similar to that of dense stochastic block models described above. We remark that each of these encodings arises directly from design matrices and tensors based on $K_{r,t}$ used in the dense Bernoulli rotation step of our reductions.

Appendix D. Further Directions and Open Problems

In this section, we describe several further directions and problems left open in this work. These directions mainly concern the PC$_\rho$ conjecture and our reduction techniques.

Further Evidence for PC$_\rho$ Conjectures. In this work, we give evidence for the PC$_\rho$ conjecture from the failure of low-degree polynomials and for specific instantiations of the PC$_\rho$ conjecture from the failure of SQ algorithms. An interesting direction for future work is to show sum of

squares lower bounds for PC_ρ and k-HPC^s supporting this conjecture. A priori, this seems to be a technically difficult task as the SOS lower bounds in Barak et al. (2016) only apply to the prior in planted clique where every vertex is included in the clique independently with probability k/n. Thus it even remains open to extend these lower bounds to the uniform prior over k-subsets of $[n]$.

How do Priors on Hidden Structure Affect Hardness? In this work, we showed that slightly altering the prior over the hidden structure of PC gave rise to a problem much more amenable to average-case reductions. This raises a broad question: for general problems $\mathcal{P}$ with hidden structure, how does changing the prior over this hidden structure affect its hardness? In other words, for natural problems other than PC, how does the conjectured computational barrier change with ρ? Another related direction for future work is whether other choices of ρ in the PC_ρ conjecture give meaningful assumptions that can be mapped to more natural problems than the ones we consider here. Furthermore, it would be interesting to study how reductions carry ensembles of problems with a general prior ρ to one another. For instance, is there a reduction between PC and another problem, such as SPCA, such that every hard prior in PC_ρ is mapped to a corresponding hard prior in SPCA?

Generalizations of Dense Bernoulli Rotations. In this work, dense Bernoulli rotations were an extremely important subroutine, serving as our simplest primitive for transforming hidden structure. An interesting technical direction for future work is to find similar transformations mapping to other distributions. More concretely, dense Bernoulli rotations approximately mapped from $\text{PB}(n, i, 1, 1/2)$ to the n distributions $\mathcal{D}_i = \mathcal{N}(c \cdot A_i, I_m)$, respectively, and mapped from $\text{Bern}(1/2)^{\otimes m}$ to $\mathcal{D} = \mathcal{N}(0, I_m)$. Are there other similar reductions mapping from these planted bit distributions to different ensembles of $\mathcal{D}, \mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$? Furthermore, can these maps be used to show tight computational lower bounds for natural problems? For example, two possibly interesting ensembles of $\mathcal{D}, \mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$ are:

1. $\mathcal{D}_i = \otimes_{j=1}^m \text{Bern}(P_{ij} n^{-\alpha})$ and some $\mathcal{D}$ where $P \in [0,1]^{n \times m}$ is a fixed matrix of constants and $\alpha > 0$.
2. $\mathcal{D}_i = \mathcal{N}(c \cdot A_i, I_m - c^2 A_i A_i^\top)$ and $\mathcal{D} = \mathcal{N}(0, I_m)$.

The first example above corresponds to whether or not there is a *sparse* analogue of Bernoulli rotations that can be used to show tight computational lower bounds. A natural approach to (1) is to apply dense Bernoulli rotations and map each entry into $\{0, 1\}$ by thresholding at some large real number $T = \Theta(\sqrt{\log n})$. While this maps to an ensemble of the form in (1), this reduction seems *lossy*, in the sense that it discards signal in the input instance, and it does not appear to show tight computational lower bounds for any natural problem. The second example above presents a set of $\mathcal{D}_i$ with the same expected covariance matrices as $\mathcal{D}$. Note that in ordinary dense Bernoulli rotations the expected covariance matrices for each i are $I_m + c^2 \cdot A_i A_i^\top$ and often a degree-2 polynomial suffices to distinguish them from $\mathcal{D}$. More generally, a natural question is: are there analogues of dense Bernoulli rotations that are tight to algorithms given by polynomials of degree higher than 2?

General Reductions to Supervised Problems. Our last open problem is more concrete than the previous two. In our reductions to MSLR and RSLR, we crucially use a subroutine mapping to NEG-SPCA. This subroutine requires that $k = \tilde{o}(n^{1/6})$ in order to show convergence in KL divergence between the Wishart and inverse Wishart distributions. Is there a reduction that relaxes this requirement to $k = \tilde{o}(n^\alpha)$ where $1/6 < \alpha < 1/2$? Providing a reduction for α arbitrarily close to

$1/2$ would essentially fill out all parameter regimes of interest in our computational lower bounds for MSLR and RSLR. Any reduction relaxing this constraint to some α with $\alpha > 1/6$ seems as though it would require new techniques and be technically interesting. Another question related to our reductions to MSLR and RSLR is: can our label generation technique be generalized to handle more general link functions σ i.e. generalized linear models where each sample-label pair (X, y) is satisfies $y = \sigma(\langle \beta, X \rangle) + \mathcal{N}(0, 1)$? In particular, is there a reduction mapping to the canonical formulation of sparse phase retrieval with $\sigma(t) = t^2$? Although the statistical-computational gap for this formulation of sparse phase retrieval seems closely related to our computational lower bound for MSLR, any such reduction seems as though it would be interesting from a technical viewpoint.

Part II
Average-Case Reduction Techniques

Appendix E. Preliminaries and Problem Formulations

In this section, we establish notation and some preliminary observations for proving our main theorems from Section 3. We already defined our notion of computational lower bounds and solving detection and recovery problems in Section 3. In this section, we begin by stating our conventions for detection problems and adversaries. In Section E.2, we introduce the framework for reductions in total variation to show computational lower bounds for detection problems. In Section E.3, we then state detection formulations for each of our problems of interest that it will suffice to exhibit reductions to. Finally, in Section E.4, we introduce the key notation that will be used throughout the paper. Later in Section P, we discuss how our reductions and lower bounds for the detection formulations in Section E.3 imply lower bounds for natural estimation and recovery variants of our problems.

E.1. Conventions for Detection Problems and Adversaries

We begin by describing our general setup for detection problems and the notions of robustness and types adversaries that we consider.

Detection Problems. In a detection task $\mathcal{P}$, the algorithm is given a set of observations and tasked with distinguishing between two hypotheses:

- a *uniform* hypothesis H_0 corresponding to the natural noise distribution for the problem; and
- a *planted* hypothesis H_1, under which observations are generated from this distribution but with a latent planted structure.

Both H_0 and H_1 can either be simple hypotheses consisting of a single distribution or a composite hypothesis consisting of multiple distributions. Our problems typically are such that either: (1) both H_0 and H_1 are simple hypotheses; or (2) both H_0 and H_1 are composite hypotheses consisting of the set of distributions that can be induced by some constrained adversary.

As discussed in Brennan et al. (2018) and Hajek et al. (2015), when detection problems need not be composite by definition, average-case reductions to natural simple vs. simple hypothesis testing formulations are stronger and technically more difficult. In these cases, composite hypotheses typically arise because a reduction gadget precludes mapping to the natural simple vs. simple hypothesis testing formulation. We remark that simple vs. simple formulations are the hypothesis testing problems that correspond to average-case decision problems $(L, \mathcal{D})$ as in Levin's theory of average-case complexity. A survey of average-case complexity can be found in Bogdanov and Trevisan (2006a).

Adversaries. The robust estimation literature contains a number of adversaries capturing different notions of model misspecification. We consider the following three central classes of adversaries:

1. **ϵ-corruption**: A set of samples $(X_1, X_2, \ldots, X_n)$ is an ϵ-corrupted sample from a distribution $\mathcal{D}$ if they can be generated by giving a set of n samples drawn i.i.d. from $\mathcal{D}$ to an adversary who then changes at most ϵn of them arbitrarily.

2. **Huber's contamination model**: A set of samples $(X_1, X_2, \ldots, X_n)$ is an ϵ-contamination of $\mathcal{D}$ in Huber's model if

$$X_1, X_2, \ldots, X_n \sim_{\text{i.i.d.}} \text{MIX}_\epsilon(\mathcal{D}, \mathcal{D}_O)$$

where $\mathcal{D}_O$ is an unknown outlier distribution chosen by an adversary. Here, $\text{MIX}_\epsilon(\mathcal{D}, \mathcal{D}_O)$ denotes the ϵ-mixture distribution formed by sampling $\mathcal{D}$ with probability $(1-\epsilon)$ and $\mathcal{D}_O$ with probability ϵ.

3. **Semirandom adversaries**: Suppose that $\mathcal{D}$ is a distribution over collections of observations $\{X_i\}_{i\in I}$ such that an unknown subset $P \subseteq I$ of indices correspond to a planted structure. A sample $\{X_i\}_{i\in I}$ is semirandom if it can be generated by giving a sample from $\mathcal{D}$ to an adversary who is allowed decrease X_i for any $i \in I \backslash P$. Some formulations of semirandom adversaries in the literature also permit increases in X_i for any $i \in P$. Our lower bounds apply to both adversarial setups.

All adversaries in these models of robustness are computationally unbounded and have access to randomness – meaning that they also have access to any hidden structure in a problem that can be recovered information theoretically. Given a single distribution $\mathcal{D}$ over a set X, any one of these three adversaries produces a set of distributions $\text{ADV}(\mathcal{D})$ that can be obtained after corruption. When formulated as detection problems, the hypotheses H_0 and H_1 are of the form $\text{ADV}(\mathcal{D})$ for some $\mathcal{D}$. We remark that ϵ-corruption can simulate contamination in Huber's model at a slightly smaller ϵ' within $o(1)$ total variation. This is because a sample from Huber's model has $\text{Bin}(n, \epsilon')$ samples from $\mathcal{D}_O$. An adversary resampling $\min\{\text{Bin}(n, \epsilon'), \epsilon n\}$ samples from $\mathcal{D}_O$ therefore simulates Huber's model within a total variation distance bounded by standard concentration for the Binomial distribution.

E.2. Reductions in Total Variation and Computational Lower Bounds

In this section, we introduce our framework for reductions in total variation, state a general condition for deducing computational lower bounds from reductions in total variation and state a number of properties of total variation that we will use in analyzing our reductions.

Average-Case Reductions in Total Variation. We give approximate reductions in total variation to show that lower bounds for one hypothesis testing problem imply lower bounds for another. These reductions yield an exact correspondence between the asymptotic Type I+II errors of the two problems. This is formalized in the following lemma, which is Lemma 3.1 from Brennan et al. (2018) stated in terms of composite hypotheses H_0 and H_1. The main quantity in the statement of the lemma can be interpreted as the smallest total variation distance between the reduced object $\mathcal{A}(X)$ and the closest mixture of distributions from either H_0' or H_1'. The proof of this lemma is short and follows from the definition of total variation. Given a hypothesis H_i, we let $\Delta(H_i)$ denote the set of all priors over the set of distributions valid under H_i.

Lemma 14 (Lemma 3.1 in Brennan et al. (2018)) *Let $\mathcal{P}$ and $\mathcal{P}'$ be detection problems with hypotheses H_0, H_1 and H_0', H_1', respectively. Let X be an instance of $\mathcal{P}$ and let Y be an instance of $\mathcal{P}'$. Suppose there is a polynomial time computable map $\mathcal{A}$ satisfying*

$$\sup_{P\in H_0} \inf_{\pi\in\Delta(H_0')} d_{TV}\left(\mathcal{L}_P(\mathcal{A}(X)), \mathbb{E}_{P'\sim\pi}\mathcal{L}_{P'}(Y)\right) + \sup_{P\in H_1} \inf_{\pi\in\Delta(H_1')} d_{TV}\left(\mathcal{L}_P(\mathcal{A}(X)), \mathbb{E}_{P'\sim\pi}\mathcal{L}_{P'}(Y)\right) \leq \delta$$

If there is a randomized polynomial time algorithm solving $\mathcal{P}'$ with Type I+II error at most ϵ, then there is a randomized polynomial time algorithm solving $\mathcal{P}$ with Type I+II error at most $\epsilon + \delta$.

If $\delta = o(1)$, then given a blackbox solver $\mathcal{B}$ for $\mathcal{P}'_D$, the algorithm that applies $\mathcal{A}$ and then $\mathcal{B}$ solves $\mathcal{P}_D$ and requires only a single query to the blackbox. We now outline the computational model and conventions we adopt throughout this paper. An algorithm that runs in randomized polynomial time refers to one that has access to $\text{poly}(n)$ independent random bits and must run in $\text{poly}(n)$ time where n is the size of the instance of the problem. For clarity of exposition, in our reductions we assume that explicit real-valued expressions can be exactly computed and that we can sample a biased random bit $\text{Bern}(p)$ in polynomial time. We also assume that the sampling and density oracles described in Definition 24 can be computed in $\text{poly}(n)$ time. For simplicity of exposition, we assume that we can sample $\mathcal{N}(0,1)$ in $\text{poly}(n)$ time.

Deducing Strong Computational Lower Bounds for Detection from Reductions. Throughout Part III, we will use the guarantees for our reductions to show computational lower bounds. For clarity and to avoid redundancy, we will outline a general recipe for showing these hardness results. All lower bounds that will be shown in Part III are *computational lower bounds* in the sense introduced in the beginning of Section E.1. Consider a problem $\mathcal{P}$ with parameters $(n, a_1, a_2, \ldots, a_t)$ and hypotheses H_0 and H_1 with a conjectured computationally hard regime captured by the constraint set $\mathcal{C}$. In order to show a computational lower bound at $\mathcal{C}$ based on one of our hardness assumptions, it suffices to show that the following is true:

Condition E.1 (Computational Lower Bounds from Reductions) For all sequences of parameters satisfying the lower bound constraints $\{(n, a_1(n), a_2(n), \ldots, a_t(n))\}_{n=1}^{\infty} \subseteq \mathcal{C}$, there are:

1. another sequence of parameters $\{(n_i, a'_1(n_i), a'_2(n_i), \ldots, a'_t(n_i))\}_{i=1}^{\infty} \subseteq \mathcal{C}$ such that

$$\lim_{i\to\infty} \frac{\log a'_k(n_i)}{\log a_k(n_i)} = 1$$

2. a sequence of instances $\{G_i\}_{i=1}^{\infty}$ of a problem PC_ρ with hypotheses H'_0 and H'_1 that cannot be solved in polynomial time according to Conjecture 3; and

3. a polynomial time reduction $\mathcal{R}$ such that if $\mathcal{P}(n_i, a'_1(n_i), a'_2(n_i), \ldots, a'_t(n_i))$ has an instance denoted by X_i, then

$$d_{\text{TV}}\left(\mathcal{R}(G_i|H'_0), \mathcal{L}(X_i|H_0)\right) = o_{n_i}(1) \quad \text{and} \quad d_{\text{TV}}\left(\mathcal{R}(G_i|H'_1), \mathcal{L}(X_i|H_1)\right) = o_{n_i}(1)$$

This can be seen to suffice as follows. Suppose that $\mathcal{A}$ solves $\mathcal{P}$ for some possible growth rate in $\mathcal{C}$ i.e. there is a sequence $\{(n_i, a'_1(n_i), a'_2(n_i), \ldots, a'_t(n_i))\}_{i=1}^{\infty} \subseteq \mathcal{C}$ with this growth rate such that $\mathcal{A}$ has Type I+II error $1 - \Omega_{n_i}(1)$ on $\mathcal{P}(n_i, a'_1(n_i), a'_2(n_i), \ldots, a'_t(n_i))$. By Lemma 14, it follows that $\mathcal{A}\circ\mathcal{R}$ also has Type I+II error $1-\Omega_{n_i}(1)$ on the sequence of inputs $\{G_i\}_{i=1}^{\infty}$, which contradicts the conjecture that they are hard instances. The three conditions above will be verified in a number of theorems in Part III.

Remarks on Deducing Computational Lower Bounds. We make several important remarks on the recipe outlined above. In all of our applications of Condition E.1, the second sequence of parameters $(n_i, a_1'(n_i), a_2'(n_i), \ldots, a_t'(n_i))$ will either be exactly a subsequence of the original parameter sequence $(n, a_1(n), a_2(n), \ldots, a_t(n))$ or will have one parameter $a_i' \neq a_i$ different from the original. However, the ability to pass to a subsequence will be crucial in a number of cases where number-theoretic constraints on parameters impact the tightness of our computational lower bounds. These constraints will arise in our reductions to robust sparse mean estimation, robust SLR and dense stochastic block models. They are discussed more in Section L.

Properties of Total Variation. The analysis of our reductions will make use of the following well-known facts and inequalities concerning total variation distance.

Fact 15 *The distance d_{TV} satisfies the following properties:*

1. *(Tensorization) Let $P_1, P_2, \ldots, P_n$ and $Q_1, Q_2, \ldots, Q_n$ be distributions on a measurable space $(\mathcal{X}, \mathcal{B})$. Then*
$$d_{TV}\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) \leq \sum_{i=1}^n d_{TV}\left(P_i, Q_i\right)$$
2. *(Conditioning on an Event) For any distribution P on a measurable space $(\mathcal{X}, \mathcal{B})$ and event $A \in \mathcal{B}$, it holds that*
$$d_{TV}\left(P(\cdot|A), P\right) = 1 - P(A)$$
3. *(Conditioning on a Random Variable) For any two pairs of random variables (X, Y) and (X', Y') each taking values in a measurable space $(\mathcal{X}, \mathcal{B})$, it holds that*
$$d_{TV}\left(\mathcal{L}(X), \mathcal{L}(X')\right) \leq d_{TV}\left(\mathcal{L}(Y), \mathcal{L}(Y')\right) + \mathbb{E}_{y \sim Y}\left[d_{TV}\left(\mathcal{L}(X|Y=y), \mathcal{L}(X'|Y'=y)\right)\right]$$
where we define $d_{TV}\left(\mathcal{L}(X|Y=y), \mathcal{L}(X'|Y'=y)\right) = 1$ for all $y \notin \text{supp}(Y')$.

Given an algorithm $\mathcal{A}$ and distribution $\mathcal{P}$ on inputs, let $\mathcal{A}(\mathcal{P})$ denote the distribution of $\mathcal{A}(X)$ induced by $X \sim \mathcal{P}$. If $\mathcal{A}$ has k steps, let $\mathcal{A}_i$ denote the ith step of $\mathcal{A}$ and $\mathcal{A}_{i\text{-}j}$ denote the procedure formed by steps i through j. Each time this notation is used, we clarify the intended initial and final variables when $\mathcal{A}_i$ and $\mathcal{A}_{i\text{-}j}$ are viewed as Markov kernels. The next lemma from Brennan et al. (2019a) encapsulates the structure of all of our analyses of average-case reductions. Its proof is simple and included in Appendix Q.1 for completeness.

Lemma 16 (Lemma 4.2 in Brennan et al. (2019a)) *Let $\mathcal{A}$ be an algorithm that can be written as $\mathcal{A} = \mathcal{A}_m \circ \mathcal{A}_{m-1} \circ \cdots \circ \mathcal{A}_1$ for a sequence of steps $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$. Suppose that the probability distributions $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_m$ are such that $d_{TV}(\mathcal{A}_i(\mathcal{P}_{i-1}), \mathcal{P}_i) \leq \epsilon_i$ for each $1 \leq i \leq m$. Then it follows that*
$$d_{TV}\left(\mathcal{A}(\mathcal{P}_0), \mathcal{P}_m\right) \leq \sum_{i=1}^m \epsilon_i$$

The next lemma bounds the total variation between unplanted and planted samples from binomial distributions. This will serve as a key computation in the proof of correctness for the reduction primitive TO-k-PARTITE-SUBMATRIX. We remark that the total variation upper bound in this

lemma is tight in the following sense. When all of the P_i are the same, the expected value of the sum of the coordinates of the first distribution is $k(P_i - Q)$ higher than that of the second. The standard deviation of the second sum is $\sqrt{kmQ(1-Q)}$ and thus when $k(P_i - Q)^2 \gg mQ(1-Q)$, the total variation below tends to one. The proof of this lemma can be found in Appendix Q.1.

Lemma 17 *If $k, m \in \mathbb{N}$, $P_1, P_2, \ldots, P_k \in [0,1]$ and $Q \in (0,1)$, then*

$$d_{TV}\left(\otimes_{i=1}^{k} \left(\text{Bern}(P_i) + \text{Bin}(m-1, Q)\right), \text{Bin}(m, Q)^{\otimes k}\right) \leq \sqrt{\sum_{i=1}^{k} \frac{(P_i - Q)^2}{2mQ(1-Q)}}$$

Here, $\mathcal{L}_1 + \mathcal{L}_2$ denotes the convolution of two given probability measures $\mathcal{L}_1$ and $\mathcal{L}_2$. The next lemma bounds the total variation between two binomial distributions. Its proof can be found in Appendix Q.1.

Lemma 18 *Given $P \in [0,1]$, $Q \in (0,1)$ and $n \in \mathbb{N}$, it follows that*

$$d_{TV}\left(\text{Bin}(n, P), \text{Bin}(n, Q)\right) \leq |P - Q| \cdot \sqrt{\frac{n}{2Q(1-Q)}}$$

E.3. Problem Formulations as Detection Tasks

In this section, we formulate each problem for which we will show computational lower bounds as a detection problem. More precisely, for each problem $\mathcal{P}$ introduced in Section 3, we introduce a detection variant $\mathcal{P}'$ such that a blackbox for $\mathcal{P}$ also solves $\mathcal{P}'$. Some of these formulations were already implicitly introduced or will be reintroduced in future sections. We gather all of these formulations here for convenience. Throughout this work, to simplify notation, we will refer to problems $\mathcal{P}$ and their detection formulations $\mathcal{P}'$ introduced in this section using the same notation. Furthermore, we will often denote the distribution over instances under the alternative hypothesis H_1 of the detection formulation for $\mathcal{P}$ with the notation $\mathcal{P}_D$, when H_1 is a simple hypothesis. We will also often parameterize $\mathcal{P}_D$ by θ to denote $\mathcal{P}_D$ conditioned on the latent hidden structure θ. When H_1 is composite, $\mathcal{P}_D$ denotes the set of distributions permitted under H_1. These general conventions are introduced on a per problem basis in this section. In Section P, we show that our reductions and lower bounds for these detection formulations also imply lower bounds for analogous estimation and recovery variants.

Robust Sparse Mean Estimation. Our hypothesis testing formulation for $\text{RSME}(n, k, d, \tau, \epsilon)$ has hypotheses given by

$$H_0 : (X_1, X_2, \ldots, X_n) \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$$
$$H_1 : (X_1, X_2, \ldots, X_n) \sim_{\text{i.i.d.}} \text{MIX}_\epsilon\left(\mathcal{N}(\tau \cdot \mu_R, I_d), \mathcal{D}_O\right)$$

where $\mathcal{D}_O$ is any adversarially chosen outlier distribution on $\mathbb{R}^d$, where $\mu_R \in \mathbb{R}^d$ is a random k-sparse unit vector chosen uniformly at random from all such vectors with entries in $\{0, 1/\sqrt{k}\}$. Note that H_1 is a composite hypothesis here since $\mathcal{D}_O$ is arbitrary. Note also that this is a formulation of RSME in Huber's contamination model, and therefore lower bounds for this detection problem imply corresponding lower bounds under stronger ϵ-corruption adversaries.

As discussed in Section B.3, RSME is only information-theoretically feasible when $\tau = \Omega(\epsilon)$. Consider any algorithm that produces some estimate $\hat{\mu}$ satisfying that $\|\hat{\mu} - \mu\|_2 < \tau/2$ with probability $1/2+\Omega(1)$ in the estimation formulation for RSME with hidden k-sparse vector μ, as described in Section B.3. This algorithm would necessarily output some $\hat{\mu}$ with $\|\hat{\mu}\|_2 < \tau/2$ under H_0 and some $\hat{\mu}$ with $\|\hat{\mu}\|_2 > \tau/2$ under H_1 with probability $1/2 + \Omega(1)$ in the hypothesis testing formulation above, thus solving it in the sense of Section 3. Thus any computational lower bounds for this hypothesis testing formulation also implies a lower bound for the typical estimation formulation of RSME.

Dense Stochastic Block Models. Given a subset $C_1 \subseteq [n]$ of size n/k, let $\text{ISBM}_D(n, C_1, P_{11}, P_{12}, P_{22})$ denote the distribution on n-vertex graphs G' introduced in Section B.4 conditioned on C_1. Furthermore, let $\text{ISBM}_D(n, k, P_{11}, P_{12}, P_{22})$ denote the mixture of these distributions induced by choosing C_1 uniformly at random from the (n/k)-subsets of $[n]$. The problem $\text{ISBM}(n, k, P_{11}, P_{12}, P_{22})$ introduced in Section B.4 is already a hypothesis testing problem, with hypotheses

$$H_0 : G \sim \mathcal{G}(n, P_0) \quad \text{and} \quad H_1 : G \sim \text{ISBM}_D(n, k, P_{11}, P_{12}, P_{22})$$

where H_0 is a composite hypothesis and P_0 can vary over all edge densities in $(0, 1)$. As we will discuss at the end of this section, computational lower bounds for this hypothesis testing problem imply lower bounds for the problem of recovering the hidden community C_1.

Testing Hidden Partition Models. Let $C = (C_1, C_2, \ldots, C_r)$ and $D = (D_1, D_2, \ldots, D_r)$ be two fixed sequences, each consisting of disjoint K-subsets of $[n]$. Let $\text{GHPM}_D(n, r, C, D, \gamma)$ denote the distribution over random matrices $M \in \mathbb{R}^{n\times n}$ introduced in Section B.5 conditioned on the fixed sequences C and D. We denote the mixture over these distributions induced by choosing C and D independently and uniformly at random from all admissible such sequences as $\text{GHPM}_D(n, r, K, \gamma)$. Similarly, we let $\text{BHPM}_D(n, r, C, P_0, \gamma)$ denote the distribution over bipartite graphs G with two parts of size n, each indexed by $[n]$ with edges included independently with probability

$$\mathbb{P}[(i, j) \in E(G)] = \begin{cases} P_0 + \gamma & \text{if } i \in C_h \text{ and } j \in D_h \text{ for some } h \in [r] \\ P_0 - \frac{\gamma}{r-1} & \text{if } i \in C_{h_1} \text{ and } j \in D_{h_2} \text{ where } h_1 \neq h_2 \\ P_0 & \text{otherwise} \end{cases}$$

where $P_0, \gamma \in (0, 1)$ be such that $\gamma/r \leq P_0 \leq 1 - \gamma$. Then let $\text{BHPM}_D(n, r, K, P_0, \gamma)$ denote the mixture formed by choosing C and D randomly as in GHPM_D. The problems $\text{GHPM}(n, r, C, D, \gamma)$ and $\text{BHPM}(n, r, K, P_0, \gamma)$ are simple hypothesis testing problems given by

$$\begin{aligned} H_0 : M &\sim \mathcal{N}(0, 1)^{\otimes n\times n} \quad \text{and} \quad H_1 : M \sim \text{GHPM}_D(n, r, K, \gamma) \\ H_0 : G &\sim \mathcal{G}_B(n, n, P_0) \quad \text{and} \quad H_1 : G \sim \text{BHPM}_D(n, r, K, P_0, \gamma) \end{aligned}$$

where $\mathcal{G}_B(n, n, P_0)$ denotes the Erdős-Rényi distribution over bipartite graphs with two parts each indexed by $[n]$ and where each edge is included independently with probability P_0.

Semirandom Planted Dense Subgraph. Our hypothesis testing formulation for the problem $\text{SEMI-CR}(n, k, P_1, P_0)$ has observation $G \in \mathcal{G}_n$ and two composite hypotheses given by

$$\begin{aligned} H_0 &: G \sim \mathbb{P}_0 \quad \text{for some } \mathbb{P}_0 \in \text{ADV}(\mathcal{G}(n, P_0)) \\ H_1 &: G \sim \mathbb{P}_1 \quad \text{for some } \mathbb{P}_1 \in \text{ADV}(\mathcal{G}(n, k, P_1, P_0)) \end{aligned}$$

Here, $\text{ADV}(\mathcal{G}(n, k, P_1, P_0))$ denotes the set of distributions induced by a semirandom adversary that can only remove edges outside of the planted dense subgraph S. Similarly, the set $\text{ADV}(\mathcal{G}(n, P_0))$ corresponds to an adversary that can remove any edges from the Erdős-Rényi graph $\mathcal{G}(n, P_0)$. We will discuss at the end of this section, how computational lower bounds for this hypothesis testing formulation imply lower bounds for the problem of approximately recovering the vertex subset corresponding to the planted dense subgraph.

Negative Sparse PCA. Our hypothesis testing formulation for $\text{NEG-SPCA}(n, k, d, \theta)$ is the spiked covariance model introduced in Johnstone and Lu (2004) and used to formulate ordinary SPCA in Gao et al. (2017), Brennan et al. (2018) and Brennan and Bresler (2019). This problem has hypotheses given by

$$H_0 : (X_1, X_2, \ldots, X_n) \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$$
$$H_1 : (X_1, X_2, \ldots, X_n) \sim_{\text{i.i.d.}} \mathcal{N}\left(0, I_d - \theta v v^\top\right)$$

where $v \in \mathbb{R}^d$ is a k-sparse unit vector with entries in $\{0, 1/\sqrt{k}\}$ chosen uniformly at random.

Unsigned and Mixtures of SLRs. Given a vector $v \in \mathbb{R}^d$, let $\text{LR}_d(v)$ be the distribution of a single sample-label pair $(X, y) \in \mathbb{R}^d \times \mathbb{R}$ given by

$$y = \langle v, X \rangle + \eta \quad \text{where } X \sim \mathcal{N}(0, I_d) \text{ and } \eta \sim \mathcal{N}(0, 1) \text{ are independent}$$

Given a subset $S \subseteq [n]$, let $\text{MSLR}_D(n, S, d, \tau, 1/2)$ denote the distribution over n independent sample-label pairs $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ each distributed as

$$(X_i, y_i) \sim \text{LR}_d(\tau s_i v_S) \quad \text{where } s_i \sim_{\text{i.i.d.}} \text{Rad}$$

where $v_S = |S|^{-1/2} \cdot \mathbf{1}_S$ and Rad denotes the Rademacher distribution which is uniform over $\{-1, 1\}$. Note that this is a even mixture of sparse linear regressions with hidden unit vectors v_S and $-v_S$ and signal strength τ. Let $\text{MSLR}_D(n, k, d, \tau, 1/2)$ denote the mixture of these distributions induced by choosing S uniformly at random from all k-subsets of $[n]$. Our hypothesis testing formulation for $\text{MSLR}(n, k, d, \tau)$ has two simple hypotheses given by

$$H_0 : \{(X_i, y_i)\}_{i \in [n]} \sim \left(\mathcal{N}(0, I_d) \otimes \mathcal{N}\left(0, 1 + \tau^2\right)\right)^{\otimes n}$$
$$H_1 : \{(X_i, y_i)\}_{i \in [n]} \sim \text{MSLR}_D(n, k, d, \tau, 1/2)$$

Our hypothesis testing formulation of $\text{USLR}(n, k, d, \tau)$ is a simple derivative of this formulation obtained by replacing each observation (X_i, y_i) with $(X_i, |y_i|)$. We remark that, unlike RSME where an estimation algorithm trivially solved the hypothesis testing formulation, the hypothesis H_0 here is not an instance of MSLR corresponding to a hidden vector of zero. This is because the labels y_i under H_0 have variance $1 + \tau^2$, whereas they would have variance 1 if they were this instance of MSLR. However, this detection problem still yields hardness for the estimation variants of MSLR and USLR described in Section B.8, albeit with a slightly more involved argument. This is discussed in Section P.

Robust SLR. Our hypothesis testing formulation for $\text{RSLR}(n, k, d, \tau, \epsilon)$ has hypotheses given by

$$H_0 : \{(X_i, y_i)\}_{i\in[n]} \sim \left(\mathcal{N}(0, I_d) \otimes \mathcal{N}\left(0, 1+\tau^2\right)\right)^{\otimes n}$$
$$H_1 : \{(X_i, y_i)\}_{i\in[n]} \sim_{\text{i.i.d.}} \text{MIX}_\epsilon\left(\text{LR}_d(\tau v), \mathcal{D}_O\right)$$

where $\mathcal{D}_O$ is any adversarially chosen outlier distribution on $\mathbb{R}^d \times \mathbb{R}$, where $v \in \mathbb{R}^d$ is a random k-sparse unit vector chosen uniformly at random from all such vectors with entries in $\{0, 1/\sqrt{k}\}$. As with the other formulations of SLR, we defer discussing the implications of lower bounds in this formulation for the estimation task described in Section B.9 to Section P.

Tensor PCA. Let $\text{TPCA}^s_D(n, \theta)$ denote the distribution on order s tensors $T \in \mathbb{R}^{n^{\otimes s}}$ with dimensions all equal to n given by $T = v^{\otimes s} + G$ where $G \sim \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$ and $v \in \{-1, 1\}^n$ is chosen independently and uniformly at random. As already introduced in Section B.10, our hypothesis testing formulation for $\text{TPCA}^s(n, \theta)$ is given by

$$H_0 : T \sim \mathcal{N}(0,1)^{\otimes n^{\otimes s}} \quad \text{and} \quad H_1 : T \sim \text{TPCA}^s_D(n, \theta)$$

Unlike the other problems we consider, our reductions only show computational lower bounds for blackboxes solving this hypothesis testing problem with a low false positive probability. As we will show in Section N, this implies a lower bound for the canonical estimation formulation for tensor PCA.

Generalized Learning Sparse Mixtures. Let $\{\mathcal{P}_\mu\}_{\mu\in\mathbb{R}}$ and $\mathcal{Q}$ be distributions on an arbitrary measurable space $(\mathcal{X}, \mathcal{B})$ and let $\mathcal{D}$ be a distribution on $\mathbb{R}$. Let $\text{GLSM}_D(n, S, d, \{\mathcal{P}_\mu\}_{\mu\in\mathbb{R}}, \mathcal{Q}, \mathcal{D})$ denote the distribution over $X_1, X_2, \ldots, X_n \in \mathcal{X}^d$ introduced in Section B.11 and let the distribution $\text{GLSM}_D(n, k, d, \{\mathcal{P}_\mu\}_{\mu\in\mathbb{R}}, \mathcal{Q}, \mathcal{D})$ denote the mixture over these distributions induced by sampling S uniformly at random from the family of k-subsets of $[n]$. Our general sparse mixtures detection problem $\text{GLSM}(n, S, d, \{\mathcal{P}_\mu\}_{\mu\in\mathbb{R}}, \mathcal{Q}, \mathcal{D})$ is the following simple vs. simple hypothesis testing formulation

$$H_0 : (X_1, X_2, \ldots, X_n) \sim_{\text{i.i.d.}} \mathcal{Q}^{\otimes d} \quad \text{and}$$
$$H_1 : (X_1, X_2, \ldots, X_n) \sim \text{GLSM}_D\left(n, k, d, \{\mathcal{P}_\mu\}_{\mu\in\mathbb{R}}, \mathcal{Q}, \mathcal{D}\right)$$

Lower bounds for this formulation directly imply lower bounds for algorithms that return an estimate $\hat{S}$ of S given samples from $\text{GLSM}_D(n, S, d, \{\mathcal{P}_\mu\}_{\mu\in\mathbb{R}}, \mathcal{Q}, \mathcal{D})$ with $|\hat{S}\Delta S| < k/2$ with probability $1/2 + \Omega(1)$ for all $|S| \le k$. Note that under H_0, such an algorithm would output some set $\hat{S}$ of size less than $k/2$ and, under H_1, it would output a set of size greater than $k/2$, each with probability $1/2 + \Omega(1)$. Thus thresholding $|\hat{S}|$ at $k/2$ solves this detection formulation in the sense of Section 3.

E.4. Notation

In this section, we establish notation that will be used repeatedly throughout this paper. Some of these definitions are repeated later upon use for convenience. Let $\mathcal{L}(X)$ denote the distribution law of a random variable X and given two laws $\mathcal{L}_1$ and $\mathcal{L}_2$, let $\mathcal{L}_1 + \mathcal{L}_2$ denote $\mathcal{L}(X + Y)$ where $X \sim \mathcal{L}_1$ and $Y \sim \mathcal{L}_2$ are independent. Given a distribution $\mathcal{P}$, let $\mathcal{P}^{\otimes n}$ denote the distribution of $(X_1, X_2, \ldots, X_n)$ where the X_i are i.i.d. according to $\mathcal{P}$. Similarly, let $\mathcal{P}^{\otimes m\times n}$ denote the

distribution on $\mathbb{R}^{m\times n}$ with i.i.d. entries distributed as $\mathcal{P}$. We let $\mathbb{R}^{n^{\otimes s}}$ denote the set of all order s tensors with dimensions all n in size that contain n^s entries. The distribution $\mathcal{P}^{\otimes n^{\otimes s}}$ denotes a tensor of these dimensions with entries independently sampled from $\mathcal{P}$. We say that two parameters a and b are polynomial in one another if there is a constant $C > 0$ such that $a^{1/C} \le b \le a^C$ as $a \to \infty$. In this paper, we adopt the standard asymptotic notation $O(\cdot), \Omega(\cdot), o(\cdot), \omega(\cdot)$ and $\Theta(\cdot)$. We let $a \asymp b$, $a \lesssim b$ and $a \gtrsim b$ be shorthands for $a = \Theta(b), a = O(b)$ and $a = \Omega(b)$, respectively. In all problems that we consider, our main focus is on the polynomial order of growth at computational barriers, usually in terms of a natural parameter n. Given a natural parameter n that will usually be clear from context, we let $a = \tilde{O}(b)$ be a shorthand for $a = O\left(b \cdot (\log n)^c\right)$ for some constant $c > 0$, and define $\tilde{\Omega}(\cdot), \tilde{o}(\cdot), \tilde{\omega}(\cdot)$ and $\tilde{\Theta}(\cdot)$ analogously. Oftentimes, it will be true that b is polynomial in n, in which case n can be replaced by b in the definition above.

Given a finite or measurable set $\mathcal{X}$, let $\mathrm{Unif}[\mathcal{X}]$ denote the uniform distribution on $\mathcal{X}$. Let Rad be shorthand for $\mathrm{Unif}[\{-1, 1\}]$, corresponding to the special case of a Rademacher random variable. Let d_{TV}, d_{KL} and χ^2 denote total variation distance, KL divergence and χ^2 divergence, respectively. Let $\mathcal{N}(\mu, \Sigma)$ denote a multivariate normal random vector with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ, where Σ is a $d \times d$ positive semidefinite matrix, and let $\mathrm{Bern}(p)$ denote the Bernoulli distribution with probability p. Let $[n] = \{1, 2, \ldots, n\}$ and $\mathcal{G}_n$ be the set of simple graphs on n vertices. Let $\mathcal{G}(n, p)$ denote the Erdős-Rényi distribution over n-vertex graphs where each edge is included independently with probability p. Let $\mathcal{G}_B(m, n, p)$ denote the Erdős-Rényi distribution over $(m + n)$-vertex bipartite graphs with m left vertices, n right vertices and such that each of the mn possible edges included independently with probability p. Throughout this paper, we will refer to bipartite graphs with m left vertices and n right vertices and matrices in $\{0, 1\}^{m\times n}$ interchangeably. Let $\mathbf{1}_S$ denote the vector $v \in \mathbb{R}^n$ with $v_i = 1$ if $i \in S$ and $v_i = 0$ if $i \notin S$ where $S \subseteq [n]$. Let $\mathrm{MIX}_\epsilon(\mathcal{D}_1, \mathcal{D}_2)$ denote the ϵ-mixture distribution formed by sampling $\mathcal{D}_1$ with probability $(1 - \epsilon)$ and $\mathcal{D}_2$ with probability ϵ. Given a partition E of $[N]$ with k parts, let $\mathcal{U}_N(E)$ denote the uniform distribution over all k-subsets of $[N]$ containing exactly one element from each part of E.

Given a matrix $M \in \mathbb{R}^{n\times n}$, the matrix $M_{S,T} \in \mathbb{R}^{k\times k}$ where S, T are k-subsets of $[n]$ refers to the minor of M restricted to the row indices in S and column indices in T. Furthermore, $(M_{S,T})_{i,j} = M_{\sigma_S(i),\sigma_T(j)}$ where $\sigma_S : [k] \to S$ is the unique order-preserving bijection and σ_T is analogously defined. Given an index set I, subset $S \subseteq I$ and pair of distributions $(\mathcal{P}, \mathcal{Q})$, let $\mathcal{M}_I(S, \mathcal{P}, \mathcal{Q})$ denote the distribution of a collection of independent random variables $(X_i : i \in I)$ with $X_i \sim \mathcal{P}$ if $i \in S$ and $X_i \sim \mathcal{Q}$ if $i \notin S$. When S is a random set, this $\mathcal{M}_I(S, \mathcal{P}, \mathcal{Q})$ denotes a mixture over the randomness of S e.g. $\mathcal{M}_{[N]}(\mathcal{U}_N(E), \mathcal{P}, \mathcal{Q})$ denotes a mixture of $\mathcal{M}_{[N]}(S, \mathcal{P}, \mathcal{Q})$ over $S \sim \mathcal{U}_N(E)$. Generally, given an index set I and $|I|$ distributions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{|I|}$, let $\mathcal{M}_I(\mathcal{P}_i : i \in I)$ denote the distribution of independent random variables $(X_i : i \in I)$ with $X_i \sim \mathcal{P}_i$ for each $i \in I$. The planted Bernoulli distribution $\mathrm{PB}(n, i, p, q)$ is over $V \in \{0, 1\}^n$ with independent entries satisfying that $V_j \sim \mathrm{Bern}(q)$ unless $j = i$, in which case $V_i \sim \mathrm{Bern}(p)$. In other words, $\mathrm{PB}(n, i, p, q)$ is a shorthand for $\mathcal{M}_{[n]}\left(\{i\}, \mathrm{Bern}(p), \mathrm{Bern}(q)\right)$. Similarly, the planted dense subgraph distribution $\mathcal{G}(n, S, p, q)$ can be written as $\mathcal{M}_I\left(\binom{S}{2}, \mathrm{Bern}(p), \mathrm{Bern}(q)\right)$ where $I = \binom{[n]}{2}$.

Appendix F. Rejection Kernels and Reduction Preprocessing

In this section, we present several average-case reduction primitives that will serve as the key subroutines and preprocessing steps in our reductions. These include pre-existing subroutines from

the rejection kernels framework introduced in Brennan et al. (2018, 2019a); Brennan and Bresler (2019), such as univariate rejection kernels from binary inputs and GAUSSIANIZE. We introduce the primitive TO-k-PARTITE-SUBMATRIX, which is a generalization of TO-SUBMATRIX from Brennan et al. (2019a) that maps from the k-partite variant of planted dense subgraph to Bernoulli matrices, by filling in the missing diagonal and symmetrizing. We also introduce a new variant of rejection kernels called symmetric 3-ary rejection kernels that will be crucial in our reductions showing universality of lower bounds for sparse mixtures.

F.1. Gaussian Rejection Kernels

Rejection kernels are a framework in Brennan et al. (2018, 2019a); Brennan and Bresler (2019) for algorithmic changes of measure based on rejection sampling. Related reduction primitives for changes of measure to Gaussians and binomial random variables appeared earlier in Ma and Wu (2015) and Hajek et al. (2015). Rejection kernels mapping a pair of Bernoulli distributions to a target pair of scalar distributions were introduced in Brennan et al. (2018). These were extended to arbitrary high-dimensional target distributions and applied to obtain universality results for submatrix detection in Brennan et al. (2019a). A surprising and key feature of both of these rejection kernels is that they are not lossy in mapping one computational barrier to another. For instance, in Brennan et al. (2019a), multivariate rejection kernels were applied to increase the relative size k of the planted submatrix, faithfully mapping instances tight to the computational barrier at lower k to tight instances at higher k. This feature is also true of the scalar rejection kernels applied in Brennan et al. (2018).

In this work, we will only need a subset of prior results on rejection kernels. In this section, we give an overview of the key guarantees for Gaussian rejection kernels with binary inputs from Brennan et al. (2018) and for GAUSSIANIZE from Brennan and Bresler (2019). We will also need a new ternary input variant of rejection kernels that will be introduced in Section F.3. We begin by introducing the Gaussian rejection kernel $\text{RK}_G(\mu, B)$ which maps $B \in \{0, 1\}$ to a real valued output and is parameterized by some $0 < q < p \leq 1$. The map $\text{RK}_G(\mu, B)$ transforms two Bernoulli inputs approximately into Gaussians. Specifically, it satisfies the two Markov transition properties

$$\text{RK}_G(\mu, B) \approx \mathcal{N}(0, 1) \quad \text{if } B \sim \text{Bern}(q) \qquad \text{and} \qquad \text{RK}_G(\mu, B) \approx \mathcal{N}(\mu, 1) \quad \text{if } B \sim \text{Bern}(p)$$

where $\text{RK}_G(\mu, B)$ can be computed in $\text{poly}(n)$ time, the $\approx$ above are up to $O_n(n^{-3})$ total variation distance and $\mu = \Theta(1/\sqrt{\log n})$. The maps $\text{RK}_G(\mu, B)$ can be implemented with the rejection sampling scheme shown in Figure 3. The total variation guarantees for Gaussian rejection kernels are captured formally in the following theorem.

Lemma 19 (Gaussian Rejection Kernels – Lemma 5.4 in Brennan et al. (2018)) *Let R_{RK} be a parameter and suppose that $p = p(R_{\text{RK}})$ and $q = q(R_{\text{RK}})$ satisfy that $0 < q < p \leq 1$, $\min(q, 1 - q) = \Omega(1)$ and $p - q \geq R_{\text{RK}}^{-O(1)}$. Let $\delta = \min\left\{\log\left(\frac{p}{q}\right), \log\left(\frac{1-q}{1-p}\right)\right\}$. Suppose that $\mu = \mu(R_{\text{RK}}) \in (0, 1)$ satisfies that*

$$\mu \leq \frac{\delta}{2\sqrt{6 \log R_{\text{RK}} + 2 \log(p - q)^{-1}}}$$

Then the map RK_G with $N = \lceil 6\delta^{-1} \log R_{\text{RK}} \rceil$ iterations can be computed in $\text{poly}(R_{\text{RK}})$ time and satisfies

$$d_{TV}\left(\text{RK}_G(\mu, \text{Bern}(p)), \mathcal{N}(\mu, 1)\right) = O\left(R_{\text{RK}}^{-3}\right) \quad \text{and} \quad d_{TV}\left(\text{RK}_G(\mu, \text{Bern}(q)), \mathcal{N}(0, 1)\right) = O\left(R_{\text{RK}}^{-3}\right)$$

Algorithm $\mathrm{RK}_G(\mu, B)$

Parameters: Input $B \in \{0,1\}$, Bernoulli probabilities $0 < q < p \leq 1$, Gaussian mean μ, number of iterations N, let $\varphi_\mu(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^2\right)$ denote the density of $\mathcal{N}(\mu, 1)$

1. Initialize $z \leftarrow 0$.
2. Until z is set or N iterations have elapsed:
   (1) Sample $z' \sim \mathcal{N}(0,1)$ independently.
   (2) If $B = 0$, if the condition
   $$p \cdot \varphi_0(z') \geq q \cdot \varphi_\mu(z')$$
   holds, then set $z \leftarrow z'$ with probability $1 - \frac{q \cdot \varphi_\mu(z')}{p \cdot \varphi_0(z')}$.
   (3) If $B = 1$, if the condition
   $$(1-q) \cdot \varphi_\mu(z'+\mu) \geq (1-p) \cdot \varphi_0(z'+\mu)$$
   holds, then set $z \leftarrow z' + \mu$ with probability $1 - \frac{(1-p) \cdot \varphi_0(z'+\mu)}{(1-q) \cdot \varphi_\mu(z'+\mu)}$.
3. Output z.

Algorithm GAUSSIANIZE

Parameters: Collection of variables $X_i \in \{0,1\}$ for $i \in I$ where I is some index set with $|I| = n$, rejection kernel parameter R_{RK}, Bernoulli probabilities $0 < q < p \leq 1$ with $p - q = R_{\mathrm{RK}}^{-O(1)}$ and $\min(q, 1-q) = \Omega(1)$ and a target means $0 \leq \mu_i \leq \tau$ for each $i \in I$ where $\tau > 0$ is a parameter

1. Form the collection of variables $Y \in \mathbb{R}^I$ by setting
   $$Y_i \leftarrow \mathrm{RK}_G(\mu_i, X_i)$$
   for each $i \in I$ where each RK_G is run with parameter R_{RK} and $N_{\mathrm{it}} = \lceil 6\delta^{-1} \log R_{\mathrm{RK}} \rceil$ iterations where $\delta = \min\left\{\log\left(\frac{p}{q}\right), \log\left(\frac{1-q}{1-p}\right)\right\}$.
2. Output the collection of variables $(Y_i : i \in I)$.

Figure 3: Gaussian instantiation of the rejection kernel algorithm from Brennan et al. (2018) and the reduction GAUSSIANIZE for mapping from Bernoulli to Gaussian planted problems from Brennan and Bresler (2019).

The proof of this lemma consists of showing that the distributions of the outputs $\mathrm{RK}_G(\mu, \mathrm{Bern}(p))$ and $\mathrm{RK}_G(\mu, \mathrm{Bern}(q))$ are close to $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(0,1)$ when conditioned to lie in the set of x with $\frac{1-p}{1-q} \leq \frac{\varphi_\mu(x)}{\varphi_0(x)} \leq \frac{p}{q}$ and then showing that this event occurs with probability close to one. The original framework in Brennan et al. (2018) mapped binary inputs to more general pairs of target

distributions than $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(0, 1)$, however we will only require binary-input rejection kernels in the Gaussian. A multivariate extension of this framework appeared in Brennan et al. (2019a).

Given an index set I, subset $S \subseteq I$ and pair of distributions $(\mathcal{P}, \mathcal{Q})$, let $\mathcal{M}_I(S, \mathcal{P}, \mathcal{Q})$ denote the distribution of a collection of independent random variables $(X_i : i \in I)$ with $X_i \sim \mathcal{P}$ if $i \in S$ and $X_i \sim \mathcal{Q}$ if $i \notin S$. More generally, given an index set I and $|I|$ distributions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{|I|}$, let $\mathcal{M}_I(\mathcal{P}_i : i \in I)$ denote the distribution of independent random variables $(X_i : i \in I)$ with $X_i \sim \mathcal{P}_i$ for each $i \in I$. For example, a planted clique in $\mathcal{G}(n, 1/2)$ on the set $S \subseteq [n]$ can be written as $\mathcal{M}_I\left(\binom{S}{2}, \text{Bern}(1), \text{Bern}(1/2)\right)$ where $I = \binom{[n]}{2}$.

We now review the guarantees for the subroutine GAUSSIANIZE. The variant presented here is restated from Brennan and Bresler (2019) to be over a general index set I rather than matrices, and with the rejection kernel parameter R_{RK} decoupled from the size n of I, as shown in Figure 3. GAUSSIANIZE maps a set of planted Bernoulli random variables to a set of independent Gaussian random variables with corresponding planted means. The procedure applies a Gaussian rejection kernel entrywise and its total variation guarantees follow by a simple application of the tensorization property of d_{TV} from Fact 15.

Lemma 20 (Gaussianization – Lemma 4.5 in Brennan and Bresler (2019)) *Let I be an index set with $|I| = n$ and let R_{RK}, $0 < q < p \leq 1$ and δ be as in Lemma 19. Let μ_i be such that $0 \leq \mu_i \leq \tau$ for each $i \in I$ where the parameter $\tau > 0$ satisfies that*

$$\tau \leq \frac{\delta}{2\sqrt{6 \log R_{\text{RK}} + 2 \log(P - Q)^{-1}}}$$

The algorithm $\mathcal{A} =$ GAUSSIANIZE runs in $\text{poly}(n, R_{\text{RK}})$ time and satisfies that

$$d_{TV}\left(\mathcal{A}(\mathcal{M}_I(S, \text{Bern}(P), \text{Bern}(Q))),\ \mathcal{M}_I\left(\mathcal{N}(\mu_i \cdot \mathbf{1}(i \in S), 1) : i \in I\right)\right) = O\left(n \cdot R_{\text{RK}}^{-3}\right)$$

for all subsets $S \subseteq I$.

F.2. Cloning and Planting Diagonals

We begin by reviewing the subroutine GRAPH-CLONE, shown in Figure 4, which was introduced in Brennan et al. (2019a) and produces several independent samples from a planted subgraph problem given a single sample. Its properties as a Markov kernel are stated in the next lemma, which is proven by showing the two explicit expressions for $\mathbb{P}[x^{ij} = v]$ in Step 1 define valid probability distributions and then explicitly writing the mass functions of $\mathcal{A}\left(\mathcal{G}(n, q)\right)$ and $\mathcal{A}\left(\mathcal{G}(n, S, p, q)\right)$.

Lemma 21 (Graph Cloning – Lemma 5.2 in Brennan et al. (2019a)) *Let $t \in \mathbb{N}$, $0 < q < p \leq 1$ and $0 < Q < P \leq 1$ satisfy that*

$$\frac{1-p}{1-q} \leq \left(\frac{1-P}{1-Q}\right)^t \quad \text{and} \quad \left(\frac{P}{Q}\right)^t \leq \frac{p}{q}$$

Then the algorithm $\mathcal{A} =$ GRAPH-CLONE runs in $\text{poly}(t, n)$ time and satisfies that for each $S \subseteq [n]$,

$$\mathcal{A}\left(\mathcal{G}(n, q)\right) \sim \mathcal{G}(n, Q)^{\otimes t} \quad \text{and} \quad \mathcal{A}\left(\mathcal{G}(n, S, p, q)\right) \sim \mathcal{G}(n, S, P, Q)^{\otimes t}$$

Algorithm GRAPH-CLONE

Inputs: Graph $G \in \mathcal{G}_n$, the number of copies t, parameters $0 < q < p \leq 1$ and $0 < Q < P \leq 1$ satisfying $\frac{1-p}{1-q} \leq \left(\frac{1-P}{1-Q}\right)^t$ and $\left(\frac{P}{Q}\right)^t \leq \frac{p}{q}$

1. Generate $x^{ij} \in \{0,1\}^t$ for each $1 \leq i < j \leq n$ such that:
   - If $\{i,j\} \in E(G)$, sample x^{ij} from the distribution on $\{0,1\}^t$ with
   $$\mathbb{P}[x^{ij} = v] = \frac{1}{p-q}\left[(1-q)\cdot P^{|v|_1}(1-P)^{t-|v|_1} - (1-p)\cdot Q^{|v|_1}(1-Q)^{t-|v|_1}\right]$$
   - If $\{i,j\} \notin E(G)$, sample x^{ij} from the distribution on $\{0,1\}^t$ with
   $$\mathbb{P}[x^{ij} = v] = \frac{1}{p-q}\left[p\cdot Q^{|v|_1}(1-Q)^{t-|v|_1} - q\cdot P^{|v|_1}(1-P)^{t-|v|_1}\right]$$
2. Output the graphs $(G_1, G_2, \ldots, G_t)$ where $\{i,j\} \in E(G_k)$ if and only if $x_k^{ij} = 1$.

Figure 4: Subroutine GRAPH-CLONE for producing independent samples from planted graph problems from Brennan et al. (2019a).

Graph cloning more generally produces a method to clone a set of Bernoulli random variables indexed by a general index set I instead of the possible edges of a graph on the vertex set $[n]$. The guarantees for this subroutine are stated in the following lemma. We remark that both of these lemmas will always be applied with $t = O(1)$, resulting in a constant loss in signal strength.

Lemma 22 (Bernoulli Cloning) *Let I be an index set with $|I| = n$, let $t \in \mathbb{N}$, $0 < q < p \leq 1$ and $0 < Q < P \leq 1$ satisfy that*

$$\frac{1-p}{1-q} \leq \left(\frac{1-P}{1-Q}\right)^t \quad \text{and} \quad \left(\frac{P}{Q}\right)^t \leq \frac{p}{q}$$

There is an algorithm $\mathcal{A} =$ BERNOULLI-CLONE that runs in $\mathrm{poly}(t,n)$ time and satisfying

$$\mathcal{A}\left(\mathcal{M}_I(\mathrm{Bern}(q))\right) \sim \mathcal{M}_I(\mathrm{Bern}(Q))^{\otimes t} \quad \text{and}$$
$$\mathcal{A}\left(\mathcal{M}_I(S, \mathrm{Bern}(p), \mathrm{Bern}(q))\right) \sim \mathcal{M}_I(S, \mathrm{Bern}(P), \mathrm{Bern}(Q))^{\otimes t}$$

for each $S \subseteq I$.

We now introduce the procedure TO-k-PARTITE-SUBMATRIX, which is shown in Figure 5 and will be crucial in our reductions to dense variants of the stochastic block model. This reduction clones the upper half of the adjacency matrix of the input graph problem to produce an independent lower half and plants diagonal entries while randomly embedding into a larger matrix to hide the diagonal entries in total variation. TO-k-PARTITE-SUBMATRIX is similar to TO-SUBMATRIX

Algorithm TO-k-PARTITE-SUBMATRIX

Inputs: k-PDS instance $G \in \mathcal{G}_N$ with clique size k that divides N and partition E of $[N]$, edge probabilities $0 < q < p \leq 1$ with $q = N^{-O(1)}$ and target dimension $n \geq \left(\frac{p}{Q} + 1\right) N$ where $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}} \left(\sqrt{q} - 1\right)$ and k divides n

1. Apply GRAPH-CLONE to G with edge probabilities $P = p$ and $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}} \left(\sqrt{q} - 1\right)$ and $t = 2$ clones to obtain (G_1, G_2).

2. Let F be a partition of $[n]$ with $[n] = F_1 \cup F_2 \cup \cdots \cup F_k$ and $|F_i| = n/k$. Form the matrix $M_{\text{PD}} \in \{0,1\}^{n \times n}$ as follows:

   (1) For each $t \in [k]$, sample $s_1^t \sim \text{Bin}(N/k, p)$ and $s_2^t \sim \text{Bin}(n/k, Q)$ and let S_t be a subset of F_t with $|S_t| = N/k$ selected uniformly at random. Sample $T_1^t \subseteq S_t$ and $T_2^t \subseteq F_t \backslash S_t$ with $|T_1^t| = s_1^t$ and $|T_2^t| = \max\{s_2^t - s_1^t, 0\}$ uniformly at random.

   (2) Now form the matrix M_{PD} such that its (i,j)th entry is

$$(M_{\text{PD}})_{ij} = \begin{cases} \mathbf{1}_{\{\pi_t(i), \pi_t(j)\} \in E(G_1)} & \text{if } i < j \text{ and } i, j \in S_t \\ \mathbf{1}_{\{\pi_t(i), \pi_t(j)\} \in E(G_2)} & \text{if } i > j \text{ and } i, j \in S_t \\ \mathbf{1}_{\{i \in T_1^t\}} & \text{if } i = j \text{ and } i, j \in S_t \\ \mathbf{1}_{\{i \in T_2^t\}} & \text{if } i = j \text{ and } i, j \in F_t \backslash S_t \\ \sim_{\text{i.i.d.}} \text{Bern}(Q) & \text{if } i \neq j \text{ and } (i,j) \notin S_t^2 \text{ for a } t \in [k] \end{cases}$$

   where $\pi_t : S_t \to E_t$ is a bijection chosen uniformly at random.

3. Output the matrix M_{PD} and the partition F.

Figure 5: Subroutine TO-k-PARTITE-SUBMATRIX for mapping from an instance of k-partite planted dense subgraph to a k-partite Bernoulli submatrix problem.

in Brennan et al. (2019a) and TO-BERNOULLI-SUBMATRIX in Brennan and Bresler (2019) but ensures that the random embedding step accounts for the k-partite promise of the input k-PDS instance. Completing the missing diagonal entries in the adjacency matrix will be crucial to apply one of our main techniques, Bernoulli rotations, which will be introduced in the next section.

The next lemma states the total variation guarantees of TO-k-PARTITE-SUBMATRIX and is a k-partite variant of Theorem 6.1 in Brennan et al. (2019a). Although technically more subtle than the analysis of TO-SUBMATRIX in Brennan et al. (2019a), this proof is tangential to our main reduction techniques and deferred to Appendix Q.2. Given a partition E of $[N]$ with k parts, let $\mathcal{U}_N(E)$ denote the uniform distribution over k-subsets of $[N]$ containing exactly one element from each part of E.

Lemma 23 (Reduction to k-Partite Bernoulli Submatrix Problems) *Let $0 < q < p \leq 1$ and $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}} \left(\sqrt{q} - 1\right)$. Suppose that n and N are such that*

$$n \geq \left(\frac{p}{Q} + 1\right) N \quad \textit{and} \quad k \leq QN/4$$

Also suppose that $q = N^{-O(1)}$ *and both* N *and* n *are divisible by* k*. Let* $E = (E_1, E_2, \dots, E_k)$ *and* $F = (F_1, F_2, \dots, F_k)$ *be partitions of* $[N]$ *and* $[n]$*, respectively. Then it follows that the algorithm* $\mathcal{A} = \text{TO-}k\text{-PARTITE-SUBMATRIX}$ *runs in* $\text{poly}(N)$ *time and satisfies*

$$d_{TV}\left(\mathcal{A}(\mathcal{G}(N, \mathcal{U}_N(E), p, q)),\ \mathcal{M}_{[n]\times[n]}\left(\mathcal{U}_n(F), \text{Bern}(p), \text{Bern}(Q)\right)\right) \le 4k \cdot \exp\left(-\frac{Q^2N^2}{48pkn}\right) + \sqrt{\frac{C_Q k^2}{2n}}$$

$$d_{TV}\left(\mathcal{A}(\mathcal{G}(N, q)),\ \text{Bern}\left(Q\right)^{\otimes n\times n}\right) \le 4k \cdot \exp\left(-\frac{Q^2N^2}{48pkn}\right)$$

where $C_Q = \max\left\{\frac{Q}{1-Q}, \frac{1-Q}{Q}\right\}$.

For completeness, we give an intuitive summary of the technical subtleties arising in the proof of this lemma. After applying GRAPH-CLONE, the adjacency matrix of the input graph G is still missing its diagonal entries. The main difficulty in producing these diagonal entries is to ensure that entries corresponding to vertices in the planted subgraph are properly sampled from $\text{Bern}(p)$. To do this, we randomly embed the original $N \times N$ adjacency matrix in a larger $n \times n$ matrix with i.i.d. entries from $\text{Bern}(Q)$ and sample all diagonal entries corresponding to entries of the original matrix from $\text{Bern}(p)$. The diagonal entries in the new $n - N$ columns are chosen so that the supports on the diagonals within each F_t each have size $\text{Bin}(n/k, Q)$. Even though this causes the sizes of the supports on the diagonals in each F_t to have the same distribution under both H_0 and H_1, the randomness of the embedding and the fact that $k = o(\sqrt{n})$ ensures that this is hidden in total variation.

F.3. Symmetric 3-ary Rejection Kernels

In this section, we introduce symmetric 3-ary rejection kernels, which will be the key gadget in our reduction showing universality of lower bounds for learning sparse mixtures in Section O. In order to map to universal formulations of sparse mixtures, it is crucial to produce a nontrivial instance of a sparse mixture with multiple planted distributions. Since previous rejection kernels all begin with binary inputs, they do not have enough degrees of freedom to map to three output distributions. The symmetric 3-ary rejection kernels 3-SRK introduced in this section overcome this issue by mapping from distributions supported on $\{-1, 0, 1\}$ to three output distributions $\mathcal{P}_+, \mathcal{P}_-$ and $\mathcal{Q}$. In order to produce clean total variation guarantees, these rejection kernels also exploit symmetry in their three input distributions on $\{-1, 0, 1\}$.

Let $\text{Tern}(a, \mu_1, \mu_2)$ where $a \in (0, 1)$ and $\mu_1, \mu_2 \in \mathbb{R}$ denote the probability distribution on $\{-1, 0, 1\}$ such that if $B \sim \text{Tern}(a, \mu_1, \mu_2)$ then

$$\mathbb{P}[X = -1] = \frac{1-a}{2} - \mu_1 + \mu_2, \quad \mathbb{P}[X = 0] = a - 2\mu_2, \quad \mathbb{P}[X = 1] = \frac{1-a}{2} + \mu_1 + \mu_2$$

if all three of these probabilities are nonnegative. The map $3\text{-SRK}(B)$, shown in Figure 6, sends an input $B \in \{-1, 0, 1\}$ to a set X simultaneously satisfying three Markov transition properties:

1. if $B \sim \text{Tern}(a, \mu_1, \mu_2)$, then $3\text{-SRK}(B)$ is close to $\mathcal{P}_+$ in total variation;

Algorithm 3-SRK$(B, \mathcal{P}_+, \mathcal{P}_-, \mathcal{Q})$

Parameters: Input $B \in \{-1, 0, 1\}$, number of iterations N, parameters $a \in (0, 1)$ and sufficiently small nonzero $\mu_1, \mu_2 \in \mathbb{R}$, distributions $\mathcal{P}_+, \mathcal{P}_-$ and $\mathcal{Q}$ over a measurable space $(X, \mathcal{B})$ such that $(\mathcal{P}_+, \mathcal{Q})$ and $(\mathcal{P}_-, \mathcal{Q})$ are computable pairs

1. Initialize z arbitrarily in the support of $\mathcal{Q}$.
2. Until z is set or N iterations have elapsed:
   (1) Sample $z' \sim \mathcal{Q}$ independently and compute the two quantities
   $$\mathcal{L}_1(z') = \frac{d\mathcal{P}_+}{d\mathcal{Q}}(z') - \frac{d\mathcal{P}_-}{d\mathcal{Q}}(z') \quad \text{and} \quad \mathcal{L}_2(z') = \frac{d\mathcal{P}_+}{d\mathcal{Q}}(z') + \frac{d\mathcal{P}_-}{d\mathcal{Q}}(z') - 2$$
   (2) Proceed to the next iteration if it does not hold that
   $$2|\mu_1| \geq |\mathcal{L}_1(z')| \quad \text{and} \quad \frac{2|\mu_2|}{\max\{a, 1-a\}} \geq |\mathcal{L}_2(z')|$$
   (3) Set $z \leftarrow z'$ with probability $P_A(x, B)$ where
   $$P_A(x, B) = \frac{1}{2} \cdot \begin{cases} 1 + \frac{a}{4\mu_2} \cdot \mathcal{L}_2(z') + \frac{1}{4\mu_1} \cdot \mathcal{L}_1(z') & \text{if } B = 1 \\ 1 - \frac{1-a}{4\mu_2} \cdot \mathcal{L}_2(z') & \text{if } B = 0 \\ 1 + \frac{1}{4\mu_2} \cdot \mathcal{L}_2(z') - \frac{a}{4\mu_1} \cdot \mathcal{L}_1(z') & \text{if } B = -1 \end{cases}$$
3. Output z.

Figure 6: 3-ary symmetric rejection kernel algorithm.

2. if $B \sim \text{Tern}(a, -\mu_1, \mu_2)$, then 3-SRK$(B)$ is close to $\mathcal{Q}$ in total variation; and
3. if $B \sim \text{Tern}(a, 0, 0)$, then 3-SRK$(B)$ is close to $\mathcal{P}_-$ in total variation.

In order to state our main results for 3-SRK(B), we will need the notion of computable pairs from Brennan et al. (2019a). The definition below is that given in Brennan et al. (2019a), without the assumption of finiteness of KL divergences. This assumption was convenient for the Chernoff exponent analysis needed for multivariate rejection kernels in Brennan et al. (2019a). Since our rejection kernels are univariate, we will be able to state our universality conditions directly in terms of tail bounds rather than Chernoff exponents.

Definition 24 (Relaxed Computable Pair Brennan et al. (2019a)) *Define a pair of sequences of distributions $(\mathcal{P}, \mathcal{Q})$ over a measurable space $(X, \mathcal{B})$ where $\mathcal{P} = (\mathcal{P}_n)$ and $\mathcal{Q} = (\mathcal{Q}_n)$ to be computable if:*

1. *there is an oracle producing a sample from $\mathcal{Q}_n$ in* poly(n) *time;*

2. *for all n, $\mathcal{P}_n$ and $\mathcal{Q}_n$ are mutually absolutely continuous and the likelihood ratio satisfies*

$$\mathbb{E}_{x\sim\mathcal{Q}_n}\left[\frac{d\mathcal{P}_n}{d\mathcal{Q}_n}(x)\right] = \mathbb{E}_{x\sim\mathcal{P}_n}\left[\left(\frac{d\mathcal{P}_n}{d\mathcal{Q}_n}(x)\right)^{-1}\right] = 1$$

   where $\frac{d\mathcal{P}_n}{d\mathcal{Q}_n}$ is the Radon-Nikodym derivative; and

3. *there is an oracle computing $\frac{d\mathcal{P}_n}{d\mathcal{Q}_n}(x)$ in $\mathrm{poly}(n)$ time for each $x \in X$.*

We remark that the second condition above always holds for discrete distributions and generally for most well-behaved distributions $\mathcal{P}$ and $\mathcal{Q}$. We now state our main total variation guarantees for 3-SRK. The proof of the next lemma follows a similar structure to the analysis of rejection sampling as in Lemma 5.1 of Brennan et al. (2018) and Lemma 5.1 of Brennan et al. (2019a). However, the bounds that we obtain are different than those in Brennan et al. (2018, 2019a) because of the symmetry of the three input Tern distributions. The proof of this lemma is deferred to Appendix Q.3.

Lemma 25 (Symmetric 3-ary Rejection Kernels) *Let $a \in (0,1)$ and $\mu_1, \mu_2 \in \mathbb{R}$ be nonzero and such that $\mathrm{Tern}(a,\mu_1,\mu_2)$ is well-defined. Let $\mathcal{P}_+, \mathcal{P}_-$ and $\mathcal{Q}$ be distributions over a measurable space $(X,\mathcal{B})$ such that $(\mathcal{P}_+,\mathcal{Q})$ and $(\mathcal{P}_-,\mathcal{Q})$ are computable pairs with respect to a parameter n. Let $S \subseteq X$ be the set*

$$S = \left\{x \in X : 2|\mu_1| \geq \left|\frac{d\mathcal{P}_+}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_-}{d\mathcal{Q}}(x)\right| \quad \text{and} \quad \frac{2|\mu_2|}{\max\{a,1-a\}} \geq \left|\frac{d\mathcal{P}_+}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_-}{d\mathcal{Q}}(x) - 2\right|\right\}$$

Given a positive integer N, then the algorithm 3-SRK : $\{-1,0,1\} \to X$ can be computed in $\mathrm{poly}(n,N)$ time and satisfies that

$$\left.\begin{matrix} d_{TV}\left(3\text{-SRK}(\mathrm{Tern}(a,\mu_1,\mu_2)),\mathcal{P}_+\right) \\ d_{TV}\left(3\text{-SRK}(\mathrm{Tern}(a,-\mu_1,\mu_2)),\mathcal{P}_-\right) \\ d_{TV}\left(3\text{-SRK}(\mathrm{Tern}(a,0,0)),\mathcal{Q}\right)\end{matrix}\right\} \leq 2\delta\left(1+|\mu_1|^{-1}+|\mu_2|^{-1}\right)+\left(\frac{1}{2}+\delta\left(1+|\mu_1|^{-1}+|\mu_2|^{-1}\right)\right)^N$$

where $\delta > 0$ is such that $\mathbb{P}_{X\sim\mathcal{P}_+}[X \notin S]$, $\mathbb{P}_{X\sim\mathcal{P}_-}[X \notin S]$ and $\mathbb{P}_{X\sim\mathcal{Q}}[X \notin S]$ are upper bounded by δ.

Appendix G. Dense Bernoulli Rotations

In this section, we formally introduce dense Bernoulli rotations and constructions for their design matrices and tensors, which will play an essential role in all of our reductions. For an overview of the main high level ideas underlying these techniques, see Sections C.2 and C.3. As mentioned in Sections C.2, dense Bernoulli rotations map $\mathrm{PB}(T,i,p,q)$ to $\mathcal{N}\left(\mu\lambda^{-1}\cdot A_i, I_m\right)$ for each $i \in [T]$ and $\mathrm{Bern}(q)^{\otimes T}$ to $\mathcal{N}(0,I_m)$ approximately in total variation, where $\mu = \tilde{\Theta}(1)$, the vectors $A_1, A_2, \ldots, A_T \in \mathbb{R}^m$ are for us to design and λ is an upper bound on the singular values of the matrix with columns A_i.

Simplifying some technical details, our reduction to RSME in Section I.1 roughly proceeds as follows: (1) its input is a k-BPC instance with parts of size M and N and biclique dimensions $k = k_M$ and k_N; (2) it applies dense Bernoulli rotations with $p = 1$ and $q = 1/2$ to the Mk_N

vectors of length $T = N/k_N$ representing the adjacency patterns in $\{0,1\}^{N/k_N}$ between each of the M left vertices and each part in the partition of the right vertices; and (3) it pads the resulting matrix with standard normals so that it has d rows. Under H_1, the result is a $d \times k_N m$ matrix $\mathbf{1}_S u^\top + \mathcal{N}(0,1)^{\otimes d \times k_N m}$ where S is the left vertex set of the biclique and u consists of scaled concatenations of the A_i. We design the adversary so that the target data matrix D in RSME is roughly of the form

$$D_{ij} \sim \begin{cases} \mathcal{N}\left(\tau k^{-1/2}, 1\right) & \text{if } i \in S \text{ and } j \text{ is not corrupted} \\ \mathcal{N}\left(\epsilon^{-1}(1-\epsilon)\tau k^{-1/2}, 1\right) & \text{if } i \in S \text{ and } j \text{ is corrupted} \\ \mathcal{N}(0,1) & \text{otherwise} \end{cases}$$

for each $i \in [d]$ and $j \in [n]$ where $n = k_N m$. Matching the two distributions above, we arrive at the following desiderata for the A_i.

- We would like each $\lambda^{-1} A_i$ to consist of $(1-\epsilon')m$ entries equal to $\tau k^{-1/2}$ and $\epsilon' m$ entries equal to $\epsilon'^{-1}(1-\epsilon')\tau k^{-1/2}$ where τ is just below the desired computational barrier $\tau = \tilde{\Theta}(k^{1/2}\epsilon^{1/2} n^{-1/4})$ and $\epsilon' \le \epsilon$ where $\epsilon' = \Theta(\epsilon)$.
- Now observe that the norm of any such $\lambda^{-1} A_i$ is $\Theta\left(\tau \epsilon^{-1/2} m^{1/2} k^{-1/2}\right)$ which is just below a norm of $\tilde{\Theta}(m^{1/2} n^{-1/4})$ at the computational barrier for RSME. Note that the normalization by λ^{-1} ensures that each $\lambda^{-1} A_i$ has ℓ_2 norm at most 1. To be as close to the computational barrier as possible, it is necessary that $m^{1/2} n^{-1/4} = \tilde{\Theta}(1)$ which rearranges to $m = \tilde{\Theta}(k_N)$ since $n = k_N m$.
- When the input is an instance of k-BPC nearly at its computational barrier, we have that $N = \tilde{\Theta}(k_N^2)$ and thus our necessary condition above implies that $m = \tilde{\Theta}(N/k_N) = \tilde{\Theta}(T)$, and hence that A is nearly square. Furthermore, if we take the A_i to be unit vectors, our desiderata that the $\lambda^{-1} A_i$ have norm $\tilde{\Theta}(m^{1/2} n^{-1/4})$ reduces to $\lambda = \tilde{\Theta}(1)$.

Summarizing this discussion, we arrive at exactly the three conditions outlined in Section C.3. We remark that while these desiderata are tailored to RSME, they will also turn out to be related to the desired properties of A in our other reductions. We now formally introduce dense Bernoulli rotations.

G.1. Mapping Planted Bits to Spiked Gaussian Tensors

Let $\mathrm{PB}(n, i, p, q)$ and $\mathrm{PB}(S, i, p, q)$ denote the planted bit distributions defined in Sections C.2 and E.4. The procedures BERN-ROTATIONS and its derivative TENSOR-BERN-ROTATIONS are shown in Figure 7. Recall that the subroutine GAUSSIANIZE was introduced in Figure 3. Note that positive semidefinite square roots of $n \times n$ matrices can be computed in $\mathrm{poly}(n)$ time. The two key Markov transition properties for these procedures that will be used throughout the paper are as follows.

Lemma 26 (Dense Bernoulli Rotations) *Let m and n be positive integers and let $A \in \mathbb{R}^{m \times n}$ be a matrix with singular values all at most $\lambda > 0$. Let R_{RK}, $0 < q < p \le 1$ and μ be as in Lemma 19. Let $\mathcal{A}$ denote* BERN-ROTATIONS *applied with rejection kernel parameter R_{RK}, Bernoulli probability parameters $0 < q < p \le 1$, output dimension m, matrix A with singular value upper bound λ and*

Algorithm BERN-ROTATIONS

Inputs: Vector $V \in \{0,1\}^n$, rejection kernel parameter R_{RK}, Bernoulli probability parameters $0 < q < p \leq 1$, output dimension m, an $m \times n$ matrix A with singular values all at most $\lambda > 0$, intermediate mean parameter $\mu > 0$

1. Form $V_1 \in \{0,1\}^n$ by applying GAUSSIANIZE to the entries in the vector V with rejection kernel parameter R_{RK}, Bernoulli probabilities q and p and target mean parameters all equal to μ.
2. Sample a vector $U \sim \mathcal{N}(0,1)^{\otimes m}$ and let $\left(I_m - \lambda^{-2} \cdot AA^\top\right)^{1/2}$ be the positive semidefinite square root of $I_m - \lambda^{-2} \cdot AA^\top$. Now form the vector
$$V_2 = \lambda^{-1} \cdot AV_1 + \left(I_m - \lambda^{-2} \cdot AA^\top\right)^{1/2} U$$
3. Output the vector V_2.

Algorithm TENSOR-BERN-ROTATIONS

Inputs: Order s tensor $T \in \mathcal{T}_{s,n}(\{0,1\})$, rejection kernel parameter R_{RK}, Bernoulli probability parameters $0 < q < p \leq 1$, output dimension m, an $m \times n$ matrices $A_1, A_2, \ldots, A_s$ with singular values less than or equal to $\lambda_1, \lambda_2, \ldots, \lambda_s > 0$, respectively, mean parameter $\mu > 0$

1. Flatten T into the vector $V_1 \in \{0,1\}^{n^s}$, form the Kronecker product $A = A_1 \otimes A_2 \otimes \cdots \otimes A_s$ and set $\lambda = \lambda_1 \lambda_2 \cdots \lambda_s$.
2. Let V_2 be the output of BERN-ROTATIONS applied to V_1 with parameters R_{RK}, $0 < q < p \leq 1, A, \lambda, \mu$ and output dimension m^s.
3. Rearrange the entries of V_2 into a tensor $T_1 \in \mathcal{T}_{s,m}(\mathbb{R})$ and output T_1.

Figure 7: Subroutines BERN-ROTATIONS and TENSOR-BERN-ROTATIONS for producing spiked Gaussian vectors and tensors, respectively, from the planted bits distribution.

mean parameter μ. Then $\mathcal{A}$ runs in $\text{poly}(n, R_{\text{RK}})$ *time and it holds that*

$$d_{TV}\left(\mathcal{A}\left(\text{PB}(n,i,p,q)\right), \mathcal{N}\left(\mu\lambda^{-1} \cdot A_{\cdot,i}, I_m\right)\right) = O\left(n \cdot R_{\text{RK}}^{-3}\right)$$
$$d_{TV}\left(\mathcal{A}\left(\text{Bern}(q)^{\otimes n}\right), \mathcal{N}\left(0, I_m\right)\right) = O\left(n \cdot R_{\text{RK}}^{-3}\right)$$

for all $i \in [n]$, where $A_{\cdot,i}$ denotes the ith column of A.

Proof Let $\mathcal{A}_1$ denote the first step of $\mathcal{A} =$ BERN-ROTATIONS with input V and output V_1, and let $\mathcal{A}_2$ denote the second step of $\mathcal{A}$ with input V_1 and output V_2. Fix some index $i \in [n]$. Now Lemma 20 implies

$$d_{\text{TV}}\left(\mathcal{A}_1\left(\text{PB}(n,i,p,q)\right), \mathcal{N}\left(\mu \cdot e_i, I_n\right)\right) = O\left(n \cdot R_{\text{RK}}^{-3}\right) \tag{1}$$

where $e_i \in \mathbb{R}^n$ is the ith canonical basis vector. Suppose that $V_1 \sim \mathcal{N}(\mu \cdot e_i, I_n)$ and let $V_1 = \mu \cdot e_i + W$ where $W \sim \mathcal{N}(0, I_n)$. Note that the entries of AW are jointly Gaussian and $\mathrm{Cov}(AW) = AA^\top$. Therefore, we have that

$$AV_1 = \mu \cdot A_{\cdot,i} + AW \sim \mathcal{N}\left(\mu \cdot A_{\cdot,i}, AA^\top\right)$$

If $U \sim \mathcal{N}(0,1)^{\otimes m}$ is independent of W, then the entries of $AW + \left(\lambda^2 \cdot I_m - \cdot AA^\top\right)^{1/2} U$ are jointly Gaussian. Furthermore, since both terms are mean zero and independent the covariance matrix of this vector is given by

$$\begin{aligned}\mathrm{Cov}\left(AW + \left(\lambda^2 \cdot I_m - AA^\top\right)^{1/2} U\right) &= \mathrm{Cov}\left(AW\right) + \mathrm{Cov}\left(\left(\lambda^2 \cdot I_m - AA^\top\right)^{1/2} U\right) \\ &= AA^\top + (\lambda^2 \cdot I_m - AA^\top) = \lambda^2 \cdot I_m\end{aligned}$$

Therefore it follows that $AW + \left(\lambda^2 \cdot I_m - AA^\top\right)^{1/2} U \sim \mathcal{N}(0, \lambda^2 \cdot I_m)$ and furthermore that

$$V_2 = \lambda^{-1} \cdot AV_1 + \left(I_m - \lambda^{-2} \cdot AA^\top\right)^{1/2} U \sim \mathcal{N}\left(\mu\lambda^{-1} \cdot A_{\cdot,i}, I_m\right)$$

Where $V_2 \sim \mathcal{A}_2\left(\mathcal{N}\left(\mu \cdot e_i, I_n\right)\right)$. Now applying $\mathcal{A}_2$ to both distributions in Equation (1) and the data-processing inequality prove that $d_{\mathrm{TV}}\left(\mathcal{A}\left(\mathrm{PB}(n,i,p,q)\right), \mathcal{N}\left(\mu\lambda^{-1} \cdot A_{\cdot,i}, I_m\right)\right) = O\left(n \cdot R_{\mathrm{RK}}^{-3}\right)$. This argument analyzing $\mathcal{A}_2$ applied with $\mu = 0$ yields that $\mathcal{A}_2\left(\mathcal{N}(0, I_n)\right) \sim \mathcal{N}(0, I_m)$. Combining this with

$$d_{\mathrm{TV}}\left(\mathcal{A}_1\left(\mathrm{Bern}(q)^{\otimes n}\right), \mathcal{N}\left(0, I_n\right)\right) = O\left(n \cdot R_{\mathrm{RK}}^{-3}\right)$$

from Lemma 20 now yields the bound $d_{\mathrm{TV}}\left(\mathcal{A}\left(\mathrm{Bern}(q)^{\otimes n}\right), \mathcal{N}\left(0, I_n\right)\right) = O\left(n \cdot R_{\mathrm{RK}}^{-3}\right)$, which completes the proof of the lemma. ■

Corollary 27 (Tensor Bernoulli Rotations) *Let s, m and n be positive integers, let $A_1, A_2, \ldots, A_s \in \mathbb{R}^{m \times n}$ be matrices with singular values less than or equal to $\lambda_1, \lambda_2, \ldots, \lambda_s > 0$, respectively. Let R_{RK}, $0 < q < p \leq 1$ and μ be as in Lemma 19. Let $\mathcal{A}$ denote* TENSOR-BERN-ROTATIONS *applied with parameters $0 < q < p \leq 1$, output dimension m, matrix $A = A_1 \otimes A_2 \otimes \cdots \otimes A_s$ with singular value upper bound $\lambda = \lambda_1\lambda_2 \cdots \lambda_s$ and mean parameter μ. If s is a constant, then $\mathcal{A}$ runs in* $\mathrm{poly}(n, R_{\mathrm{RK}})$ *time and it holds that for each $e \in [n]^s$,*

$$\begin{aligned}d_{TV}\left(\mathcal{A}\left(\mathrm{PB}_s(n,e,p,q)\right), \mathcal{N}\left(\mu(\lambda_1\lambda_2\cdots\lambda_s)^{-1} \cdot A_{\cdot,e_1} \otimes A_{\cdot,e_2} \otimes \cdots \otimes A_{\cdot,e_s}, I_m^{\otimes s}\right)\right) &= O\left(n^s \cdot R_{\mathrm{RK}}^{-3}\right) \\ d_{TV}\left(\mathcal{A}\left(\mathrm{Bern}(q)^{\otimes n^{\otimes s}}\right), \mathcal{N}\left(0, I_m^{\otimes s}\right)\right) &= O\left(n^s \cdot R_{\mathrm{RK}}^{-3}\right)\end{aligned}$$

where $A_{\cdot,i}$ denotes the ith column of A.

Proof Let σ_i^j for $1 \leq i \leq r_j$ be the nonzero singular values of A_j for each $1 \leq j \leq s$. Then the nonzero singular values of the Kronecker product $A = A_1 \otimes A_2 \otimes \cdots \otimes A_s$ are all of the products $\sigma_{i_1}^1 \sigma_{i_2}^2 \cdots \sigma_{i_s}^s$ for all $(i_1, i_2, \ldots, i_s)$ with $1 \leq i_j \leq r_j$ for each $1 \leq j \leq s$. Thus if $\sigma_i^j \leq \lambda_j$ for each $1 \leq j \leq s$, then $\lambda = \lambda_1\lambda_2 \cdots \lambda_s$ is an upper bound on the singular values of A. The corollary now follows by applying Lemma 26 with parameters p, q, μ and λ, matrix A, output dimension m^s and input dimension n^s. ■

G.2. $\mathbb{F}_r^t$ Design Matrices

In this section, we introduce a family of matrices $K_{r,t}$ that plays a key role in constructing the matrices A in our applications of dense Bernoulli rotations. Throughout this section, r will denote a prime number and t will denote a fixed positive integer. As outlined in the beginning of this section and in Section C.3, there are three desiderata of the matrices $K_{r,t}$ that are needed for our applications of dense Bernoulli rotations. In the context of $K_{r,t}$, these three properties are:

1. The rows of $K_{r,t}$ are unit vectors and close to orthogonal in the sense that the largest singular value of $K_{r,t}$ is bounded above by a constant.
2. The matrices $K_{r,t}$ both contain exactly two distinct real values as entries.
3. The matrices $K_{r,t}$ contain a fraction of approximately $1/r$ negative entries per column.

The matrices $K_{r,t}$ are constructed based on the incidence structure of the points in $\mathbb{F}_r^t$ with the Grassmanian of hyperplanes in $\mathbb{F}_r^t$ and their affine shifts. The construction of $K_{r,t}$ is motivated by the projective geometry codes and their applications to constructing 2-block designs. We remark that a classic trick counting the number of ordered d-tuples of linearly independent vectors in $\mathbb{F}_r^t$ shows that the number of d-dimensional subspaces of $\mathbb{F}_r^t$ is

$$|\mathrm{Gr}(d, \mathbb{F}_r^t)| = \frac{(r^t-1)(r^t-r)\cdots(r^t-r^{d-1})}{(r^d-1)(r^d-r)\cdots(r^d-r^{d-1})}$$

This implies that the number of hyperplanes in $\mathbb{F}_r^t$ is $\ell = \frac{r^t-1}{r-1}$. We now give the definition of the matrix $K_{r,t}$ as a weighted incidence matrix between the points of $\mathbb{F}_r^t$ and affine shifts of the hyperplanes in the Grassmanian $\mathrm{Gr}(t-1, \mathbb{F}_r^t)$.

Definition 28 (Design Matrices $K_{r,t}$) *Let $P_1, P_2, \ldots, P_{r^t}$ be an enumeration of the points in $\mathbb{F}_r^t$ and $V_1, V_2, \ldots, V_\ell$, where $\ell = \frac{r^t-1}{r-1}$, be an enumeration of the hyperplanes in $\mathbb{F}_r^t$. For each V_i, let $u_i \neq 0$ denote a vector in $\mathbb{F}_r^t$ not contained in V_i. Define $K_{r,t} \in \mathbb{R}^{r\ell \times r^t}$ to be the matrix with the following entries*

$$(K_{r,t})_{r(i-1)+a+1,j} = \frac{1}{\sqrt{r^t(r-1)}} \cdot \begin{cases} 1 & \text{if } P_j \notin V_i + au_i \\ 1-r & \text{if } P_j \in V_i + au_i \end{cases}$$

for each $a \in \{0, 1, \ldots, r-1\}$ where $V_i + v$ denotes the affine shift of V_i by v.

We now establish the key properties of $K_{r,t}$ in the following simple lemma. Note that the lemma implies that the submatrix consisting of the rows of $K_{r,t}$ corresponding to hyperplanes in $\mathbb{F}_r^t$ has rows that are exactly orthogonal. However, the additional rows of $K_{r,t}$ corresponding to affine shifts of these hyperplanes will prove crucial in preserving *tightness to algorithms* in our average-case reductions. As established in the subsequent lemma, these additional rows only mildly perturb the largest singular value of the matrix.

Lemma 29 (Sub-orthogonality of $K_{r,t}$) *If $r \geq 2$ is prime, then $K_{r,t}$ satisfies that:*

1. *for each $1 \leq i \leq kr\ell$, it holds that $\|(K_{r,t})_i\|_2 = 1$;*

2. *the inner product between the rows $(K_{r,t})_i$ and $(K_{r,t})_j$ where $i \neq j$ are given by*

$$\langle (K_{r,t})_i, (K_{r,t})_j \rangle = \begin{cases} -(r-1)^{-1} & \text{if } \lfloor (i-1)/r \rfloor = \lfloor (j-1)/r \rfloor \\ 0 & \text{otherwise} \end{cases}$$

3. *each column of $K_{r,t}$ contains exactly $\frac{r^t-1}{r-1}$ entries equal to $\frac{1-r}{\sqrt{r^t(r-1)}}$.*

Proof Let r_i denote the ith row $(K_{r,t})_i$ of $K_{r,t}$. Fix a pair $1 \leq i < j \leq r\ell$ and let $1 \leq i' \leq j' \leq \ell$ and $a, b \in \{0, 1, \ldots, r-1\}$ be such that $i = r(i'-1) + a$ and $j = r(j'-1) + b$. The affine subspaces of $\mathbb{F}_r^t$ corresponding to r_i and r_j are then $A_i = V_{i'} + au_{i'}$ and $A_j = V_{j'} + bu_{j'}$. Observe that

$$\|r_i\|_2^2 = (r^t - |A_i|) \cdot \frac{1}{r^t(r-1)} + |A_i| \cdot \frac{(1-r)^2}{r^t(r-1)} = 1$$

Similarly, we have that

$$\langle r_i, r_j \rangle = (r^t - |A_i \cup A_j|) \cdot \frac{1}{r^t(r-1)} + (|A_i \cup A_j| - |A_i \cap A_j|) \cdot \frac{1-r}{r^t(r-1)} + |A_i \cap A_j| \cdot \frac{(1-r)^2}{r^t(r-1)}$$

for each $1 \leq i, j \leq r\ell$. Since the size of a subspace is invariant under affine shifts, we have that $|A_i| = |V_{i'}| = |A_j| = |V_{j'}| = r^{t-1}$. Furthermore, since $A_i \cap A_j$ is the intersection of two affine shifts of subspaces of dimension $t-1$ of $\mathbb{F}_r^t$, it follows that $A_i \cap A_j$ is either empty, an affine shift of a $(t-2)$-dimensional subspace or equal to both A_i and A_j. Note that if $i \neq j$, then A_i and A_j are distinct. We remark that when $t = 1$, each A_i is an affine shift of the trivial hyperplane $\{0\}$ and thus is a singleton. Now note that the intersection $A_i \cap A_j$ is only empty if A_i and A_j are affine shifts of one another which occurs if and only if $\lfloor (i-1)/r \rfloor = i' = j' = \lfloor (j-1)/r \rfloor$. In this case, it follows that $|A_i \cup A_j| = |A_i| + |A_j| = 2r^{t-1}$. In this case, we have

$$\langle r_i, r_j \rangle = (r^t - 2r^{t-1}) \cdot \frac{1}{r^t(r-1)} + 2r^{t-1} \cdot \frac{1-r}{r^t(r-1)} = -(r-1)^{-1}$$

If $i' \neq j'$, then $A_i \cap A_j$ is the affine shift of a $(t-2)$-dimensional subspace which implies that $|A_i \cap A_j| = r^{t-1}$. Furthermore, $|A_i \cup A_j| = |A_i| + |A_j| - |A_i \cap A_j| = 2r^{t-1} - r^{t-2}$. In this case, we have that

$$\langle r_i, r_j \rangle = (r-1)^2 \cdot \frac{1}{r^2(r-1)} - 2(r-1) \cdot \frac{1}{r^2} + \frac{r-1}{r^2} = 0$$

This completes the proof of (2). We remark that this last case never occurs if $t = 1$. Now note that any point is in exactly one affine shift of each V_i. Therefore each column contains exactly ℓ negative entries, which proves (3). ■

The next lemma uses the computation of $\langle (K_{r,t})_i, (K_{r,t})_j \rangle$ above to compute the singular values of $K_{r,t}$.

Lemma 30 *The nonzero singular values of $K_{r,t}$ are $\sqrt{1 + (r-1)^{-1}}$ with multiplicity $(r-1)\ell$.*

Proof Lemma 29 shows that $(K_{r,t})(K_{r,t})^\top$ is block-diagonal with ℓ blocks of dimension $r \times r$. Furthermore, each block is of the form $\left(1 + (r-1)^{-1}\right) I_r - (r-1)^{-1} \mathbf{1}\mathbf{1}^\top$. The eigenvalues of each of these blocks are $1 + (r-1)^{-1}$ with multiplicity $r-1$ and 0 with multiplicity 1. Thus the eigenvalues of $(K_{r,t})(K_{r,t})^\top$ are $1 + (r-1)^{-1}$ and 0 with multiplicities $(r-1)\ell$ and ℓ, respectively, implying the result. ■

G.3. $\mathbb{F}_r^t$ Design Tensors

In this section, we introduce a family of tensors $T_{r,t}^{(V_i,V_j,L)}$ that will be used in TENSOR-BERN-ROTATIONS in the matrix case with $s = 2$ to map to hidden partition models in Section M.2. An overview of how these tensors will be used in dense Bernoulli rotations was given in Section C.3. Similar to the previous section, the $T_{r,t}^{(V_i,V_j,L)}$ are constructed to have the following properties:

1. Given a pair of hyperplanes (V_i, V_j) and a linear function $L : \mathbb{F}_r \to \mathbb{F}_r$, the slice $T_{r,t}^{(V_i,V_j,L)}$ of the constructed tensor is an $r^t \times r^t$ matrix with Frobenius norm $\left\|T_{r,t}^{(V_i,V_j,L)}\right\|_F = 1$.
2. These slices are approximately orthogonal in the sense that the Gram matrix with entries given by the matrix inner products $\mathrm{Tr}\left(T_{r,t}^{(V_i,V_j,L)} \cdot T_{r,t}^{(V_{i'},V_{j'},L')}\right)$ has a bounded spectral norm.
3. Each slice $T_{r,t}^{(V_i,V_j,L)}$ contains two distinct entries and is an average signed adjacency matrix of a hidden partition model i.e. has these two entries arranged into an r-block community structure.
4. Matrices formed by specific concatentations of $T_{r,t}^{(V_i,V_j,L)}$ into larger matrices remain the average signed adjacency matrices of hidden partition models. This will be made precise in Lemma 36 and will be important in our reduction from k-PC.

The construction of the family of tensors $T_{r,t}^{(V_i,V_j,L)}$ is another construction using the incidence geometry of $\mathbb{F}_r^t$, but is more involved than the two constructions in the previous section. Throughout this section, we let $V_1, V_2, \ldots, V_\ell$ and $P_1, P_2, \ldots, P_{r^t}$ be an enumeration of the hyperplanes and points of $\mathbb{F}_r^t$ as in Definition 28. Furthermore, for each V_i, we fix a particular point $u_i \neq 0$ of $\mathbb{F}_r^t$ not contained in V_i. In order to introduce the family $T_{r,t}^{(V_i,V_j,L)}$, we first define the following important class of bipartite graphs.

Definition 31 (Affine Block Graphs $G_{r,t}$) *For each $1 \leq i \leq \ell$, let $A_0^i \cup A_1^i \cup \cdots \cup A_{r-1}^i$ be the partition of $\mathbb{F}_r^t$ given by the affine shifts $A_x^i = (V_i + xu_i)$ for each $x \in \mathbb{F}_r$. Given two hyperplanes V_i, V_j and linear function $L : \mathbb{F}_r \to \mathbb{F}_r$, define the bipartite graph $G_{r,t}(V_i, V_j, L)$ with two parts of size r^t, each indexed by points in $\mathbb{F}_r^t$, as follows:*

1. *all of the edges between the points with indices in A_x^i in the left part of $G_{r,t}(V_i, V_j, L)$ and the points with indices in A_y^j in the right part are present if $L(x) = y$; and*
2. *none of the edges between the points of A_x^i on the left and A_y^j on the right are present if $L(x) \neq y$.*

We now define the slices of the tensor $T_{r,t}$ to be weighted adjacency matrices of the bipartite graphs $G_{r,t}(V_i, V_j, L)$ as in the following definition.

Definition 32 (Design Tensors $T_{r,t}$) *For any two hyperplanes V_i, V_j and linear function $L : \mathbb{F}_r \to \mathbb{F}_r$, define the $r^t \times r^t$ matrix $T_{r,t}^{(V_i,V_j,L)}$ to have entries given by*

$$\left(T_{r,t}^{(V_i,V_j,L)}\right)_{k,l} = \frac{1}{r^t\sqrt{r-1}} \cdot \begin{cases} r-1 & \text{if } (P_k, P_l) \in E\left(G_{r,t}(V_i, V_j, L)\right) \\ -1 & \text{otherwise} \end{cases}$$

for each $1 \leq k, l \leq r^t$.

The next two lemmas establish that the tensor $T_{r,t}$ satisfies the four desiderata discussed above, which will be crucial in our reduction to hidden partition models.

Lemma 33 (Sub-orthogonality of $T_{r,t}$) *If $r \geq 2$ is prime, then $T_{r,t}$ satisfies that:*

1. *for each $1 \leq i, j \leq r^t$ and linear function L, it holds that $\left\|T_{r,t}^{(V_i,V_j,L)}\right\|_F = 1$;*
2. *the inner product between the slices $T_{r,t}^{(V_i,V_j,L)}$ and $T_{r,t}^{(V_{i'},V_{j'},L')}$ where $(V_i, V_j, L) \neq (V_{i'}, V_{j'}, L')$ is*

$$\mathrm{Tr}\left(T_{r,t}^{(V_i,V_j,L)} \cdot T_{r,t}^{(V_{i'},V_{j'},L')}\right) = \begin{cases} -(r-1)^{-1} & \text{if } (V_i,V_j) = (V_{i'},V_{j'}) \text{ and } L = L' + a \text{ for some } a \neq 0 \\ 0 & \text{if } (V_i,V_j) \neq (V_{i'},V_{j'}) \text{ or } L \neq L' + a \text{ for all } a \in \mathbb{F}_r \end{cases}$$

Proof Fix two triples (V_i, V_j, L) and $(V_{i'}, V_{j'}, L')$ and let $G_1 = G_{r,t}(V_i, V_j, L)$ and $G_2 = G_{r,t}(V_{i'}, V_{j'}, L')$. Now observe that

$$\begin{aligned}\mathrm{Tr}\left(T_{r,t}^{(V_i,V_j,L)} \cdot T_{r,t}^{(V_{i'},V_{j'},L')}\right) = {} & \frac{1}{r^{2t}(r-1)} \cdot (r-1)^2 \cdot |E(G_1) \cap E(G_2)| \\ & - \frac{1}{r^{2t}(r-1)} \cdot (r-1) \cdot (|E(G_1) \cup E(G_2)| - |E(G_1) \cap E(G_2)|) \\ & + \frac{1}{r^{2t}(r-1)} \cdot \left(r^{2t} - |E(G_1) \cup E(G_2)|\right) \end{aligned} \tag{2}$$

Now note that since L is a function, there are exactly r pairs $(x, y) \in \mathbb{F}_r^2$ such that $L(x) = y$ and thus exactly r pairs of left and right sets (A_x^i, A_y^j) that are completely connected by edges in G_1. This implies that there are $|E(G_1)| = |E(G_2)| = r^{2t-1}$ edges in both G_1 and G_2. We now will show that

$$|E(G_1) \cap E(G_2)| = \begin{cases} r^{2t-1} & \text{if } (V_i,V_j,L) = (V_{i'},V_{j'},L') \\ r^{2t-2} & \text{if } (V_i,V_j) \neq (V_{i'},V_{j'}) \text{ or } L \neq L' + a \text{ for all } a \in \mathbb{F}_r \\ 0 & \text{if } (V_i,V_j) = (V_{i'},V_{j'}) \text{ and } L = L' + a \text{ for some } a \neq 0 \end{cases} \tag{3}$$

We remark that, as in the proof of Lemma 29, it is never true that $(V_i, V_j) \neq (V_{i'}, V_{j'})$ if $t = 1$. The first case follows immediately from the fact that $|E(G_1)| = r^{2t-1}$. Now consider the case in which $V_i \neq V_{i'}$ and $V_j \neq V_{j'}$. As in the proof of Lemma 29, any pair of affine spaces A_x^i and $A_{x'}^{i'}$ either intersects in an affine space of dimension $t-2$, an affine space of dimension $t-1$ if $A_x^i = A_{x'}^{i'}$ are equal and in the empty set if A_x^i and $A_{x'}^{i'}$ are affine shifts of one another. Since $V_i \neq V_{i'}$, only the first of these three options is possible. Therefore, for all $x, x', y, y' \in \mathbb{F}_r$, it follows that $(A_x^i \times A_y^j) \cap (A_{x'}^{i'} \times A_{y'}^{j'}) = (A_x^i \cap A_{x'}^{i'}) \times (A_y^j \times A_{y'}^{j'})$ has size r^{2t-4} since both $A_x^i \cap A_{x'}^{i'}$ and $A_y^j \times A_{y'}^{j'}$ are affine spaces of dimension $t-2$. Now observe that

$$|E(G_1) \cap E(G_2)| = \sum_{L(x)=y} \sum_{L'(x')=y'} \left|\left(A_x^i \times A_y^j\right) \cap \left(A_{x'}^{i'} \times A_{y'}^{j'}\right)\right| = r^2 \cdot r^{2t-4} = r^{2t-2}$$

since there are exactly r pairs (x, y) with $L(x) = y$. Now suppose that $V_i = V_{i'}$ and $V_j \neq V_{j'}$. In this case, we have that $A_x^i \cap A_{x'}^{i'}$ is empty if $x \neq x'$ and otherwise has size $|A_x^i| = r^{t-1}$. Thus it follows that

$$\left|\left(A_x^i \times A_y^j\right) \cap \left(A_{x'}^{i'} \times A_{y'}^{j'}\right)\right| = \begin{cases} r^{2t-3} & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}$$

This implies that

$$|E(G_1) \cap E(G_2)| = \sum_{L(x)=y} \sum_{L'(x')=y'} \left| \left(A_x^i \times A_y^j\right) \cap \left(A_{x'}^{i'} \times A_{y'}^{j'}\right) \right| = r \cdot r^{2t-3} = r^{2t-2}$$

since for each fixed $x = x'$, there is a unique pair (y, y') with $L(x) = y$ and $L(x') = y'$. The case in which $V_i \neq V_{i'}$ and $V_j = V_{j'}$ is handled by a symmetric argument. Now suppose that $(V_i, V_j) = (V_{i'}, V_{j'})$. It follows that $(A_x^i \times A_y^j) \cap (A_{x'}^{i'} \times A_{y'}^{j'})$ has size r^{2t-2} if $x = x'$ and $y = y'$, and is empty otherwise. The formula above therefore implies that $|E(G_1) \cap E(G_2)|$ is r^{2t-2} times the number of solutions to $L(x) = L'(x)$. Since $L - L'$ is linear, the number of solutions is 0 if $L - L'$ is constant and not equal to zero, 1 if $L - L'$ is not constant or r if $L = L'$. This completes the proof of Equation (3). Now observe that $|E(G_1) \cup E(G_2)| = |E(G_1)| + |E(G_2)| - |E(G_1) \cap E(G_2)| = 2r^{2t-1} - |E(G_1) \cap E(G_2)|$. Substituting this expression for $|E(G_1) \cup E(G_2)|$ into Equation (2) yields that

$$\mathrm{Tr}\left(T_{r,t}^{(V_i,V_j,L)} \cdot T_{r,t}^{(V_{i'},V_{j'},L')}\right) = \frac{r^2}{r^{2t}(r-1)} \cdot |E(G_1) \cap E(G_2)| - \frac{1}{r-1}$$

Combining this with the different cases of Equation (3) shows part (2) of the lemma. Part (1) of the lemma follows from this computation and fact that

$$\left\| T_{r,t}^{(V_i,V_j,L)} \right\|_F^2 = \mathrm{Tr}\left(\left(T_{r,t}^{(V_i,V_j,L)}\right)^2\right)$$

This completes the proof of the lemma. ■

We now define an unfolded matrix variant of the tensor $T_{r,t}$ that will be used in our applications of TENSOR-BERN-ROTATIONS to map to hidden partition models. The row indexing in $M_{r,t}$ will be important and related to the community alignment property of $T_{r,t}$ that will be established in Lemma 36.

Definition 34 (Unfolded Matrix $M_{r,t}$) *Let $M_{r,t}$ be an $(r-1)^2\ell^2 \times r^{2t}$ matrix with entries given by*

$$(M_{r,t})_{a(r-1)\ell^2+i'(r-1)\ell+b\ell+j'+1,ir^t+j+1} = \left(T_{r,t}^{(V_{i'+1},V_{j'+1},L_{a+1,b+1})}\right)_{i,j}$$

for each $0 \le i', j' \le (r-1)\ell - 1$, $0 \le a, b \le r-2$ and $0 \le i, j \le r^t - 1$, where $L_{c,d} : \mathbb{F}_r \to \mathbb{F}_r$ denotes the linear function given by $L_{c,d}(x) = cx + d$.

The next lemma is similar to Lemma 30 and deduces the singular values of $M_{r,t}$ from Lemma 33. The proof is very similar to that of Lemma 30.

Lemma 35 (Singular Values of $M_{r,t}$) *The nonzero singular values of $M_{r,t}$ are $\sqrt{1 + (r-1)^{-1}}$ with multiplicity $(r-1)(r-2)\ell^2$ and $(r-1)^{-1/2}$ with multiplicity $(r-1)\ell^2$.*

Proof Observe that the rows of $M_{r,t}$ are formed by vectorizing the slices of $T_{r,t}$. Thus Lemma 33 implies that $(M_{r,t})(M_{r,t})^\top$ is block-diagonal with $(r-1)\ell^2$ blocks of dimension $(r-1) \times (r-1)$, where each block corresponds to slices with indices $(V_i, V_j, L_{c,d})$ where i, j and c are fixed on

over each block while d ranges over $\{1, 2, \ldots, r-1\}$. Furthermore, each block is of the form $\left(1 + (r-1)^{-1}\right) I_{r-1} - (r-1)^{-1}\mathbf{1}\mathbf{1}^\top$. The eigenvalues of each of these blocks are $1 + (r-1)^{-1}$ with multiplicity $r-2$ and $(r-1)^{-1}$ with multiplicity 1. Thus the eigenvalues of $(M_{r,t})(M_{r,t})^\top$ are $1 + (r-1)^{-1}$ and $(r-1)^{-1}$ with multiplicities $(r-1)(r-2)\ell^2$ and $(r-1)\ell^2$, respectively, which implies the result. ■

Given m^2 matrices $M^{1,1}, M^{1,2}, \ldots, M^{k,k} \in \mathbb{R}^{n\times n}$, let $\mathcal{C}\left(M^{1,1}, M^{1,2}, \ldots, M^{k,k}\right)$ denote the matrix $X \in \mathbb{R}^{kn\times kn}$ formed by concatenating the $M^{i,j}$ with

$$X_{an+b+1,cn+d+1} = M^{a+1,c+1}_{b+1,d+1} \quad \text{for all } 0 \le a, c \le k-1 \text{ and } 0 \le b, d \le n-1$$

We refer to a matrix $M \in \mathbb{R}^{n\times n}$ as a k-block matrix for some k that divides n if there are two values $x_1, x_2 \in \mathbb{R}$ and two partitions $[n] = E_1 \cup E_2 \cup \cdots \cup E_k = F_1 \cup F_2 \cup \cdots \cup F_k$ both into parts of size n/k such that

$$M_{ij} = \begin{cases} x_1 & \text{if } (i,j) \in E_h \times F_h \text{ for some } 1 \le h \le k \\ x_2 & \text{otherwise} \end{cases}$$

The next lemma shows an alignment property of different slices of $T_{r,t}$ that will be crucial in stitching together the local applications of TENSOR-BERN-ROTATIONS with $M_{r,t}$ in our reduction to hidden partition models. This lemma will use indexing the in $M_{r,t}$ and the role of linear functions L in defining the affine block graphs $G_{r,t}$.

Lemma 36 (Community Alignment in $T_{r,t}$) *Let $1 \le s_1, s_2, \ldots, s_k \le (r-1)\ell$ be arbitrary indices and*

$$M^{i,j} = T^{(V_{i'}, V_{j'}, L)}_{r,t} \quad \text{for each } 1 \le i, j \le k$$

where i' and j' are the unique $1 \le i', j' \le \ell$ such that $i' \equiv s_i \pmod{\ell}$ and $j' \equiv s_j \pmod{\ell}$ and $L(x) = ax + b$ where $a = \lceil s_i/\ell \rceil$ and $b = \lceil s_j/\ell \rceil$. Then it follows that $\mathcal{C}\left(M^{1,1}, M^{1,2}, \ldots, M^{k,k}\right)$ is an r-block matrix.

Proof Let $t_i = i'$ be the unique $1 \le i' \le \ell$ such that $i' \equiv s_i \pmod{\ell}$ and let $a_i = \lceil s_i/\ell \rceil \in \{1, 2, \ldots, r-1\}$ for each $1 \le i \le \ell$. Furthermore, let $L_{ij}(x) = a_i x + a_j$ for $1 \le i, j \le k$ and, for each $x \in \mathbb{R}$ and $1 \le i \le \ell$, let A^i_x be the affine spaces as in Definition 31. Note that since $0 < a_i < r$, it follows that each L_{ij} is a non-constant and hence invertible linear function. Given a subset $S \subseteq \mathbb{F}_r^t$ and some $s \in \mathbb{N}$, let $I(s, S)$ denote the set of indices $I(s, S) = \{s + i : P_i \in S\}$.

Now define the partition $[kr^t] = E_0 \cup E_2 \cup \cdots \cup E_{r-1}$ as follows

$$E_i = \bigcup_{j=1}^{k} I\left((j-1)r^t, A^{t_j}_{x_{ij}}\right) \quad \text{where } x_{ij} = L^{-1}_{j1}(L_{11}(i))$$

and similarly define the partition $[kr^t] = F_0 \cup F_2 \cup \cdots \cup F_{r-1}$ as follows

$$F_i = \bigcup_{j=1}^{k} I\left((j-1)r^t, A^{t_j}_{y_{ij}}\right) \quad \text{where } y_{ij} = L_{1j}(i)$$

Let $X \in \mathbb{R}^{kr^t \times kr^t}$ denote the matrix $X = \mathcal{C}\left(M^{1,1}, M^{1,2}, \ldots, M^{k,k}\right)$. We will show that

$$X_{a,b} = \frac{r-1}{r^t\sqrt{r-1}} \quad \text{if } (a,b) \in E_i \times F_i \text{ for some } 0 \le i \le r-1 \tag{4}$$

Suppose that $(a,b) \in E_i \times F_i$ and specifically that $(j_a-1)r^t + 1 \le a \le j_a r^t$ and $(j_b-1)r^t+1 \le b \le j_b r^t$ for some $1 \le j_a, j_b \le k$. The definitions of E_i and F_i imply that $z_a \in A^{t_{j_a}}_{x_{ij_a}}$ where $z_a = P_{a-(j_a-1)r^t}$ and $z_b \in A^{t_{j_b}}_{y_{ij_b}}$ where $z_b = P_{b-(j_b-1)r^t}$. Note that

$$X_{a,b} = M^{j_a,j_b}_{a-(j_a-1)r^t, b-(j_b-1)r^t}$$

by the definition of $\mathcal{C}$. Therefore by Definition 32, it suffices to show that (z_a, z_b) is an edge of the bipartite graph $G_{r,t}(V_{t_{j_a}}, V_{t_{j_b}}, L_{j_a j_b})$ for all such (a,b) to establish (4). By Definition 31, (z_a, z_b) is an edge if and only if $L_{j_a j_b}(x_{ij_a}) = y_{ij_b}$. Observe that the definitions of x_{ij_a} and y_{ij_b} yield that

$$\begin{aligned} a_{j_a} x_{ij_a} + a_1 = L_{j_a 1}(x_{ij_a}) = L_{11}(i) = a_1 \cdot i + a_1 \\ y_{ij_b} = L_{1 j_b}(i) = a_1 \cdot i + a_{j_b} \\ L_{j_a j_b}(x) = a_{j_a} x + a_{j_b} \end{aligned} \tag{5}$$

Adding $a_{j_b} - a_1$ to both sides of Equation (5) therefore yields that

$$L_{j_a j_b}(x_{ij_a}) = a_{j_a} x_{ij_a} + a_{j_b} = a_1 \cdot i + a_{j_b} = y_{ij_b}$$

which completes the proof of (4). Now note that each $M^{i,j}$ contains exactly r^{2t-1} entries equal to $(r-1)/r^t\sqrt{r-1}$ and thus X contains exactly $k^2 r^{2t-1}$ such entries. The definitions of E_i and F_i imply that they each contain exactly kr^{t-1} elements. Thus $\cup_{i=0}^{r-1} E_i \times F_i$ contains $k^2 r^{2t-1}$ elements. Therefore (4) also implies that $X_{a,b} = -1/r^t\sqrt{r-1}$ for all $(a,b) \notin \cup_{i=0}^{r-1} E_i \times F_i$. This proves that X is an r-block matrix and completes the proof of the lemma. ■

The community alignment property shown in this lemma is directly related to the indexing of rows in $M_{r,t}$. More precisely, the above lemma implies that for any subset $S \subseteq [(r-1)\ell]$, the rows of $M_{r,t}$ indexed by elements in the support of $\mathbf{1}_S \otimes \mathbf{1}_S$ can be arranged as sub-matrices of an $|S|r^t \times |S|r^t$ matrix that is an r-block matrix. This property will be crucial in our reduction from k-PC and k-PDS to hidden partition models in Section M.2.

G.4. A Random Matrix Alternative to $K_{r,t}$

In this section, we introduce the random matrix analogue $R_{n,\epsilon}$ of $K_{r,t}$ defined below. Rather than have all independent entries, $R_{n,\epsilon}$ is constrained to be symmetric. This ends up being technically convenient, as it suffices to bound the eigenvalues of $R_{n,\epsilon}$ in order to upper bound its largest singular value. This symmetry also yields a direct connection between the eigenvalues of $R_{n,\epsilon}$ and the eigenvalues of sparse random graphs, which have been studied extensively.

Definition 37 (Random Design Matrix $R_{n,\epsilon}$) *If $\epsilon \in (0, 1/2]$, let $R_{n,\epsilon} \in \mathbb{R}^{n \times n}$ denote the random symmetric matrix with independent entries sampled as follows*

$$(R_{n,\epsilon})_{ij} = (R_{n,\epsilon})_{ji} \sim \begin{cases} -\sqrt{\frac{1-\epsilon}{\epsilon n}} & \text{with prob. } \epsilon \\ \sqrt{\frac{\epsilon}{(1-\epsilon)n}} & \text{with prob. } 1-\epsilon \end{cases}$$

for all $1 \le i < j \le n$, and $(R_{n,\epsilon})_{ii} = \sqrt{\frac{\epsilon}{(1-\epsilon)n}}$ for each $1 \le i \le n$.

We now establish the key properties of the matrix $R_{n,\epsilon}$. Consider the graph G where $\{i,j\} \in E(G)$ if and only if $(R_{n,\epsilon})_{ij}$ is negative. By definition, we have that G is an ϵ-sparse Erdős-Rényi graph with $G \sim \mathcal{G}(n,\epsilon)$. Furthermore, if A is the adjacency matrix of G, a direct calculation yields that $R_{n,\epsilon}$ can be expressed as

$$R_{n,\epsilon} = \sqrt{\frac{\epsilon}{(1-\epsilon)n}} \cdot I_n + \frac{1}{\sqrt{\epsilon(1-\epsilon)n}} \cdot (\mathbb{E}[A] - A) \tag{6}$$

A line of work has given high probability upper bounds on the largest eigenvalue of $\mathbb{E}[A] - A$ in order to study concentration of sparse Erdős-Rényi graphs in the spectral norm of their adjacency matrices (Füredi and Komlós, 1981; Vu, 2005; Feige and Ofek, 2005; Lu and Peng, 2013; Bandeira and Van Handel, 2016; Le et al., 2017). As outlined in Le et al. (2017), the works Füredi and Komlós (1981), Vu (2005) and Lu and Peng (2013) apply Wigner's trace method to obtain spectral concentration results for general random matrices that, in this context, imply with high probability that

$$\|\mathbb{E}[A] - A\| = 2\sqrt{d}\,(1 + o_n(1)) \quad \text{for } d \gg (\log n)^4$$

where $d = \epsilon n$ and $\|\cdot\|$ denotes the spectral norm on $n \times n$ symmetric matrices. In Feige and Ofek (2005), Bandeira and Van Handel (2016) and Le et al. (2017), it is shown that this requirement on d can be relaxed and that it holds with high probability that

$$\|\mathbb{E}[A] - A\| = O_n(\sqrt{d}) \quad \text{for } d = \Omega_n(\log n)$$

These results, the fact that $R_{n,\epsilon}$ is symmetric and the above expression for $R_{n,\epsilon}$ in terms of A are enough to establish our main desired properties of $R_{n,\epsilon}$, which are stated formally in the following lemma.

Lemma 38 (Key Properties of $R_{n,\epsilon}$) *If $\epsilon \in (0, 1/2]$ satisfies that $\epsilon n = \omega_n(\log n)$, there is a constant $C > 0$ such that the random matrix $R_{n,\epsilon}$ satisfies the following two conditions with probability $1 - o_n(1)$:*

1. *the largest singular value of $R_{n,\epsilon}$ is at most C; and*
2. *every column of $R_{n,\epsilon}$ contains between $\epsilon n - C\sqrt{\epsilon n \log n}$ and $\epsilon n + C\sqrt{\epsilon n \log n}$ negative entries.*

Proof The number of negative entries in the ith column of $R_{n,\epsilon}$ is distributed as $\text{Bin}(n-1,\epsilon)$. A standard Chernoff bound for the binomial distribution yields that if $X \sim \text{Bin}(n-1,\epsilon)$, then

$$\mathbb{P}\left[|X - (n-1)\epsilon| \geq \delta(n-1)\epsilon\right] \leq 2\exp\left(-\frac{\delta^2(n-1)\epsilon}{3}\right)$$

for all $\delta \in (0,1)$. Setting $\delta = C'\sqrt{n^{-1}\epsilon^{-1}\log n}$ for a sufficiently large constant $C' > 0$ and taking a union bound over all columns i implies that property (2) in the lemma statement holds with probability $1 - o_n(1)$. We now apply Theorem 1.1 in Le et al. (2017) as in the first example in Section 1.4, where the graph is not modified. Since $\epsilon n = \omega_n(\log n)$, this yields with probability $1 - o_n(1)$ that

$$\|\mathbb{E}[A] - A\| \leq C''\sqrt{d}$$

for some constant $C'' > 0$, where A and d are as defined above. The decomposition of $R_{n,\epsilon}$ in Equation (6) now implies that with probability $1 - o_n(1)$

$$\|R_{n,\epsilon}\| \le \sqrt{\frac{\epsilon}{(1-\epsilon)n}} + \frac{1}{\sqrt{\epsilon(1-\epsilon)n}} \cdot C''\sqrt{d} = O_n(1)$$

since $\epsilon \in (0, 1/2]$ and $d = \epsilon n$. This establishes that property (1) holds with probability $1 - o_n(1)$. A union bound over (1) and (2) now completes the proof of the lemma. ■

While $R_{n,\epsilon}$ and $K_{r,t}$ satisfy similar conditions needed by our reductions, they also have differences that will dictate when one is used over the other. The following highlights several key points in comparing these two matrices.

- $R_{n,\epsilon}$ and $K_{r,t}$ are analogous when $n = r^t$ and $\epsilon = 1/r$. In this case, both matrices contain the same two values $1/\sqrt{r^t(r-1)}$ and $-\sqrt{(r-1)/r^t}$. The rows of $K_{r,t}$ are unit vectors and the rows of $R_{n,\epsilon}$ are approximately unit vectors – property (2) in Lemma 38 implies that the norm of each row is $1 \pm O_n(\sqrt{(\epsilon n)^{-1}\log n})$. Like $K_{r,t}$, Lemma 38 implies that $R_{n,\epsilon}$ is also approximately orthogonal with largest singular value bounded above by a constant.

- While $K_{r,t}$ has exactly a $(1/r)$-fraction of entries in each column that are negative, $R_{n,\epsilon}$ only has *approximately* an ϵ-fraction of entries in each of its columns that are negative. For some of our reductions, such as our reductions to RSME and RSLR, having approximately an ϵ-fraction of its entries equal to the negative value in Definition 37 is sufficient. In our reductions to ISBM, GHPM, BHPM and SEMI-CR, it will be important that $K_{r,t}$ contains exactly $(1/r)$-fraction of negative entries per column. The approximate guarantee of $R_{n,\epsilon}$ would correspond to only showing lower bounds against algorithms that are *adaptive* and do not need to know the sizes of the hidden communities.

- As is mentioned in Section B.4 and will be discussed in Section L, our applications of dense Bernoulli rotations with $K_{r,t}$ will generally be tight when a natural parameter n in our problems satisfies that $\sqrt{n} = \tilde{\Theta}(r^t)$. This imposes a number theoretic condition (T) on the pair (n, r), arising from the fact that t must be an integer, which generally remains a condition in the computational lower bounds we show for ISBM, GHPM and BHPM. In contrast, $R_{n,\epsilon}$ places no number-theoretic constraints on n and ϵ, which can be arbitrary, and thus the condition (T) can be removed from our computational lower bounds for RSME and RSLR. We remark that when $r = n^{o(1)}$, which often is the regime of interest in problems such as RSME, then the condition (T) is trivial and places no further constraints on (n, r) as will be shown in Lemma 80.

- $R_{n,\epsilon}$ is random while $K_{r,t}$ is fixed. In our reductions, it is often important that the same design matrix is used throughout multiple applications of dense Bernoulli rotations. Since $R_{n,\epsilon}$ is a random matrix, this requires generating a single instance of $R_{n,\epsilon}$ and using this one instance throughout our reductions. In each of our reductions, we will rejection sample $R_{n,\epsilon}$ until it satisfies the two criteria in Lemma 38 for a maximum of $O((\log n)^2)$ rounds, and then use the resulting matrix throughout all applications of dense Bernoulli rotations in that reduction. The probability bounds in Lemma 38 imply that the probability no sample from $R_{n,\epsilon}$ satisfying these criteria is found is $n^{-\omega_n(1)}$. This is a failure mode for our reductions

and contributes a negligible $n^{-\omega_n(1)}$ to the total variation distance between the output of our reductions and their target distributions.

- For some of our reductions, applying dense Bernoulli rotations with either of the two matrices $R_{n,\epsilon}$ or $K_{r,t}$ yields the same guarantees. This is the case for our reductions to MSLR, GLSM and TPCA, where $r = 2$ and the condition (T) is trivial and mapping to columns with approximately half of their entries negative is sufficient. As mentioned above, this is also the case when $r \asymp \epsilon^{-1} = n^{o(1)}$ in RSME.

- Some differences between $R_{n,\epsilon}$ and $K_{r,t}$ that are unimportant for our reductions include that $R_{n,\epsilon}$ is exactly square while $K_{r,t}$ is only approximately square and that $R_{n,\epsilon}$ is symmetric while $K_{r,t}$ is not.

For consistency, the pseudocode and analysis for all of our reductions are written with $K_{r,t}$ rather than $R_{n,\epsilon}$. Modifying our reductions to use $R_{n,\epsilon}$ is straightforward and consists of adding the rejection sampling step to sample $R_{n,\epsilon}$ discussed above. In Sections I.1, I.2 and L, we discuss in more detail how to make these modifications to our reductions to RSME and RSLR and the computational lower bounds they yield.

There are several reasons why we present our reductions with $K_{r,t}$ rather than $R_{n,\epsilon}$. The analysis of $K_{r,t}$ in Section G.2 is simple and self-contained while the analysis of $R_{n,\epsilon}$ requires fairly involved results from random matrix theory. The construction of $K_{r,t}$ naturally extends to $T_{r,t}$ while a random tensor analogue $T_{r,t}$ seems as though it would be prohibitively difficult to analyze. Reductions with $K_{r,t}$ give an explicit encoding of cliques into the hidden structure of our target problems as discussed in Section C.8, yielding slightly stronger and more explicit computational lower bounds in this sense.

Appendix H. Negatively Correlated Sparse PCA

This section is devoted to giving a reduction from bipartite planted dense subgraph to negatively correlated sparse PCA, the high level of which was outlined in Section C.5. This reduction will be used in the next section as a crucial subroutine in reductions to establish conjectured statistical-computational gaps for two supervised problems: mixtures of sparse linear regressions and robust sparse linear regression. The analysis of this reduction relies on a result on the convergence of the Wishart distribution and its inverse. This result is proven in the second half of this section.

H.1. Reducing to Negative Sparse PCA

In this section, we give our reduction BPDS-TO-NEG-SPCA from bipartite planted dense subgraph to negatively correlated sparse PCA, which is shown in Figure 8. This reduction is described with the input bipartite graph as its adjacency matrix of Bernoulli random variables. A key subroutine in this reduction is the procedure χ^2-RANDOM-ROTATION from Brennan and Bresler (2019), which is also shown in Figure 8. The lemma below provides total variation guarantees for χ^2-RANDOM-ROTATION and is adapted from Lemma 4.6 from Brennan and Bresler (2019) to be in our notation and apply to the generalized case where the input matrix M is rectangular instead of square.

This lemma can be proven with an identical argument to that in Lemma 4.6 from Brennan and Bresler (2019), with the following adjustment of parameters to the rectangular case. The first two

Algorithm χ^2-RANDOM-ROTATION

Inputs: Matrix $M \in \{0,1\}^{m \times n}$, Bernoulli probabilities $0 < q < p \leq 1$, planted subset size k_n that divides n and a parameter $\tau > 0$

1. Sample $r_1, r_2, \ldots, r_n \sim_{\text{i.i.d.}} \sqrt{\chi^2(n/k_n)}$ and truncate the r_j with $r_j \leftarrow \min\left\{r_j, 2\sqrt{n/k_n}\right\}$ for each $j \in [n]$.
2. Compute M' by applying GAUSSIANIZE to M with Bernoulli probabilities p and q, rejection kernel parameter $R_{\text{RK}} = mn$, parameter τ and target mean values $\mu_{ij} = \frac{1}{2}\tau \cdot r_j \cdot \sqrt{k_n/n}$ for each $i \in [m]$ and $j \in [n]$.
3. Sample an orthogonal matrix $R \in \mathbb{R}^{n \times n}$ from the Haar measure on the orthogonal group $\mathcal{O}_n$ and output the columns of the matrix $M'R$.

Algorithm BPDS-TO-NEG-SPCA

Inputs: Matrix $M \in \{0,1\}^{m \times n}$, Bernoulli probabilities $0 < q < p \leq 1$, planted subset size k_n that divides n and a parameter $\tau > 0$, target dimension $d \geq m$

1. Compute $X = (X_1, X_2, \ldots, X_n)$ where $X_i \in \mathbb{R}^m$ as the columns of the matrix output by χ^2-RANDOM-ROTATION applied to M with parameters p, q, k_n and τ.
2. Compute $\hat{\Sigma} = \sum_{i=1}^n X_i X_i^\top$ and let $R \in \mathbb{R}^{m \times n}$ be the top m rows of an orthogonal matrix sampled from the Haar measure on the orthogonal group $\mathcal{O}_n$ and compute the matrix
$$M' = \sqrt{n(n-m-1)} \cdot \hat{\Sigma}^{-1/2} R$$
where $\hat{\Sigma}^{-1/2}$ is the positive semidefinite square root of the inverse of $\hat{\Sigma}$.
3. Output the columns of the $d \times n$ matrix with upper left $m \times n$ submatrix M' and all remaining entries sampled i.i.d. from $\mathcal{N}(0,1)$.

Figure 8: Subroutine χ^2-RANDOM-ROTATION for random rotations to instances of sparse PCA from Brennan and Bresler (2019) and our reduction from bipartite planted dense subgraph to negative sparse PCA.

steps of χ^2-RANDOM-ROTATION maps $\mathcal{M}_{[m]\times[n]}(S \times T, p, q)$ approximately to

$$\frac{\tau}{2}\sqrt{\frac{k_n}{n}} \cdot \mathbf{1}_S u_T^\top + \mathcal{N}(0,1)^{\otimes m \times n}$$

where u_T is the vector with $(u_T)_i = r_i$ if $i \in T$ and $(u_T)_i = 0$ otherwise. The argument in Lemma 4.6 from Brennan and Bresler (2019) shows that the final step of χ^2-RANDOM-ROTATION maps this distribution approximately to

$$\frac{\tau}{2}\sqrt{\frac{k_n}{n}} \cdot \mathbf{1}_S w^\top + \mathcal{N}(0,1)^{\otimes m \times n}$$

where $w \sim \mathcal{N}(0, I_n)$. Now observe that the entries of this matrix are zero mean and jointly Gaussian. Furthermore, the columns are independent and have covariance matrix $I_m + \frac{\tau^2 k_n |S|}{4n} \cdot v_S v_S^\top$ where $v_S = |S|^{-1/2} \cdot \mathbf{1}_S$. Summarizing the result of this argument, we have the following lemma.

Lemma 39 (χ^2 Random Rotations – Adapted from Lemma 4.6 in Brennan and Bresler (2019)) *Given parameters m, n, let $0 < q < p \leq 1$ be such that $p - q = (mn)^{-O(1)}$ and $\min(q, 1-q) = \Omega(1)$, let $k_n \leq n$ be such that k_n divides n and let $\tau > 0$ be such that*

$$\tau \leq \frac{\delta}{2\sqrt{6\log(mn) + 2\log(p-q)^{-1}}} \quad \text{where} \quad \delta = \min\left\{\log\left(\frac{p}{q}\right), \log\left(\frac{1-q}{1-p}\right)\right\}$$

The algorithm $\mathcal{A} = \chi^2$-RANDOM-ROTATION runs in $\mathrm{poly}(m, n)$ time and satisfies that

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S \times T, p, q)\right), \mathcal{N}\left(0, I_m + \frac{\tau^2 k_n |S|}{4n} \cdot v_S v_S^\top\right)^{\otimes n}\right) \leq O\left((mn)^{-1}\right) + k_n (4e^{-3})^{n/2k_n}$$

$$d_{TV}\left(\mathcal{A}\left(\mathrm{Bern}(q)^{\otimes m \times n}\right), \mathcal{N}(0,1)^{\otimes m \times n}\right) = O\left((mn)^{-1}\right)$$

where $v_S = \frac{1}{\sqrt{|S|}} \cdot \mathbf{1}_S \in \mathbb{R}^m$ for all subsets $S \subseteq [m]$ and $T \subseteq [n]$ with $|T| = k_n$.

Throughout the remainder of this section, we will need to use properties of the Wishart and inverse Wishart distributions. These distributions on random matrices are defined as follows.

Definition 40 (Wishart Distribution) *Let n and d be positive integers and $\Sigma \in \mathbb{R}^{d \times d}$ be a positive semidefinite matrix. The Wishart distribution $\mathcal{W}_d(n, \Sigma)$ is the distribution of the matrix $\hat{\Sigma} = \sum_{i=1}^n X_i X_i^\top$ where $X_1, X_2, \ldots, X_n \sim_{\text{i.i.d.}} \mathcal{N}(0, \Sigma)$.*

Definition 41 (Inverse Wishart Distribution) *Let n, d and Σ be as in Definition 40. The inverted Wishart distribution $\mathcal{W}_d^{-1}(n, \Sigma)$ is the distribution of $\hat{\Sigma}^{-1}$ where $\hat{\Sigma} \sim \mathcal{W}_d(n, \Sigma)$.*

In order to analyze BPDS-TO-NEG-SPCA, we also will need the following observation from Brennan and Bresler (2019). This is a simple consequence of the fact that the distribution $\mathcal{N}(0, I_n)$ is isotropic and thus invariant under multiplication by elements of the orthogonal group $\mathcal{O}_n$.

Lemma 42 (Lemma 6.5 in Brennan and Bresler (2019)) *Suppose that $n \geq d$ and let $\Sigma \in \mathbb{R}^{d \times d}$ be a fixed positive definite matrix and let $\Sigma_e \sim \mathcal{W}_d(n, \Sigma)$. Let $R \in \mathbb{R}^{d \times n}$ be the matrix consisting of the first d rows of an $n \times n$ matrix chosen randomly and independently of Σ_e from the Haar measure $\mu_{\mathcal{O}_n}$ on $\mathcal{O}_n$. Let $(Y_1, Y_2, \ldots, Y_n)$ be the n columns of $\Sigma_e^{1/2} R$, then $Y_1, Y_2, \ldots, Y_n \sim_{\text{i.i.d.}} \mathcal{N}(0, \Sigma)$.*

We now will state and prove the main total variation guarantees for BPDS-TO-NEG-SPCA in the theorem below. The proof of the theorem below crucially relies on the upper bound in Theorem 44 on the KL divergence between Wishart matrices and their inverses. Proving this KL divergence bound is the focus of the next subsection.

Theorem 43 (Reduction to Negative Sparse PCA) *Let m, n, p, q, k_n and τ be as in Lemma 39 and suppose that $d \geq m$ and $n \gg m^3$ as $n \to \infty$. Fix any subset $S \subseteq [m]$ and let θ_S be given by*

$$\theta_S = \frac{\tau^2 k_n |S|}{4n + \tau^2 k_n |S|}$$

Then algorithm $\mathcal{A} = \text{BPDS-TO-NEG-SPCA}$ *runs in* $\text{poly}(m,n)$ *time and satisfies that*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S\times T,p,q)\right),\mathcal{N}\left(0,I_d-\theta_S v_S v_S^\top\right)^{\otimes n}\right)\le O\left(m^{3/2}n^{-1/2}\right)+k_n(4e^{-3})^{n/2k_n}$$
$$d_{TV}\left(\mathcal{A}\left(\text{Bern}(q)^{\otimes m\times n}\right),\mathcal{N}(0,1)^{\otimes d\times n}\right)=O\left(m^{3/2}n^{-1/2}\right)$$

where $v_S=\frac{1}{\sqrt{|S|}}\cdot\mathbf{1}_S\in\mathbb{R}^d$ *for all subsets* $S\subseteq[m]$ *and* $T\subseteq[n]$ *with* $|T|=k_n$.

Proof Let $\mathcal{A}_1$ denote the application of χ^2-RANDOM-ROTATION with input M and output X in Step 1 of $\mathcal{A}$. Let $\mathcal{A}_{2a}$ denote the Markov transition with input X and output $n(n-m-1)\cdot\hat{\Sigma}^{-1}$, as defined in Step 2 of $\mathcal{A}$, and let $\mathcal{A}_{2b\text{-}3}$ denote the Markov transition with input $Y=n(n-m-1)\cdot\hat{\Sigma}^{-1}$ and output Z formed by padding $Y^{1/2}R$ with i.i.d. $\mathcal{N}(0,1)$ random variables to be $d\times n$ i.e. the output of $\mathcal{A}$. Furthermore, let $\mathcal{A}_{2\text{-}3}=\mathcal{A}_{2b\text{-}3}\circ\mathcal{A}_{2a}$ denote Steps 2 and 3 with input X and output Z.

Now fix some positive semidefinite matrix $\Sigma\in\mathbb{R}^{m\times m}$ and observe that if $A=\sum_{i=1}^n Z_iZ_i^\top\sim\mathcal{W}_m(n,I_m)$ where $Z_1,Z_2,\ldots,Z_n\sim_{\text{i.i.d.}}\mathcal{N}(0,I_m)$, then it also follows that

$$\Sigma^{1/2}A\Sigma^{1/2}=\sum_{i=1}^n\left(\Sigma^{1/2}Z_i\right)\left(\Sigma^{1/2}Z_i\right)^\top\sim\mathcal{W}_m(n,\Sigma)$$

since $\Sigma^{1/2}Z_i\sim\mathcal{N}(0,\Sigma)$. Now observe that $(\Sigma^{1/2}A\Sigma^{1/2})^{-1}=\Sigma^{-1/2}A^{-1}\Sigma^{-1/2}$ and thus if $B\sim\mathcal{W}_m^{-1}(n,I_m)$ then $\Sigma^{-1/2}B\Sigma^{-1/2}\sim\mathcal{W}_m^{-1}(n,\Sigma)$. Let $\beta^{-1}=n(n-m-1)$ and $C\sim\mathcal{W}_m^{-1}(n,\beta\cdot I_m)$. Therefore we have by the data processing inequality for total variation in Fact 15 that

$$\begin{aligned}d_{\text{TV}}\left(\mathcal{W}_m(n,\Sigma),\,\mathcal{W}_m^{-1}\left(n,\beta\cdot\Sigma^{-1}\right)\right)&=d_{\text{TV}}\left(\mathcal{L}\left(\Sigma^{1/2}A\Sigma^{1/2}\right),\mathcal{L}\left(\Sigma^{1/2}C\Sigma^{1/2}\right)\right)\\&\le d_{\text{TV}}\left(\mathcal{L}\left(A\right),\mathcal{L}\left(C\right)\right)\\&\le\sqrt{\frac{1}{2}\cdot d_{\text{KL}}\left(\mathcal{W}_m(n,I_m)\,\middle\|\,\mathcal{W}_m^{-1}(n,\beta\cdot I_m)\right)}\\&=O\left(m^{3/2}n^{-1/2}\right)\end{aligned}$$

where the last inequality follows from the fact that $n\gg m^3$, Theorem 44 and Pinsker's inequality.

Suppose that $X\sim\mathcal{N}\left(0,I_m+\theta'_Sv_Sv_S^\top\right)^{\otimes n}$ where $\theta'_S=\frac{\tau^2k_n|S|}{4n}$. Then we have that the output Y of $\mathcal{A}_{2a}$ satisfies $Y=n(n-m-1)\cdot\hat{\Sigma}^{-1}\sim\mathcal{W}_m^{-1}\left(n,\beta\cdot\Sigma^{-1}\right)$ where

$$\Sigma=\left(I_m+\theta'_Sv_Sv_S^\top\right)^{-1}=I_m-\frac{\theta'_S}{1+\theta'_S}\cdot v_S^\top v_S^\top=I_m-\theta_Sv_Sv_S^\top$$

Therefore it follows from the inequality above that

$$d_{\text{TV}}\left(\mathcal{A}_{2a}\left(\mathcal{N}\left(0,I_m+\theta'_Sv_Sv_S^\top\right)^{\otimes n}\right),\mathcal{W}_m\left(n,I_m-\theta_Sv_Sv_S^\top\right)\right)=O\left(m^{3/2}n^{-1/2}\right)$$

Similarly, if $X\sim\mathcal{N}\left(0,I_m\right)^{\otimes n}$ then we have that

$$d_{\text{TV}}\left(\mathcal{A}_{2a}\left(\mathcal{N}\left(0,I_m\right)^{\otimes n}\right),\mathcal{W}_m\left(n,I_m\right)\right)=O\left(m^{3/2}n^{-1/2}\right)$$

applying the same argument with $\Sigma = I_m$. Now note that if $Y \sim \mathcal{W}_m\left(n, I_m - \theta_S v_S v_S^\top\right)$ then Lemma 42 implies that $\mathcal{A}_{\text{2b-3}}$ produces $Z \sim \mathcal{N}\left(0, I_d - \theta_S v_S v_S^\top\right)^{\otimes n}$. Similarly, it follows that if $Y \sim \mathcal{W}_m(n, I_m)$ then Lemma 42 implies that $Z \sim \mathcal{N}\left(0, I_d\right)^{\otimes n}$.

We now will use Lemma 16 applied to the steps $\mathcal{A}_i$ above and the following sequence of distributions

$$\begin{aligned}
\mathcal{P}_0 &= \mathcal{M}_{[m]\times[n]}(S \times T, p, q) \\
\mathcal{P}_1 &= \mathcal{N}\left(0, I_m + \theta'_S v_S v_S^\top\right)^{\otimes n} \\
\mathcal{P}_{\text{2a}} &= \mathcal{W}_m\left(n, I_m - \theta_S v_S v_S^\top\right) \\
\mathcal{P}_{\text{2b-3}} &= \mathcal{N}\left(0, I_d - \theta_S v_S v_S^\top\right)^{\otimes n}
\end{aligned}$$

As in the statement of Lemma 16, let ϵ_i be any real numbers satisfying $d_{\text{TV}}\left(\mathcal{A}_i(\mathcal{P}_{i-1}), \mathcal{P}_i\right) \le \epsilon_i$ for each step i. A direct application of Lemma 39, shows that we can take $\epsilon_1 = O(m^{-1}n^{-1}) + k(4e^{-3})^{n/2k}$. The arguments above show we can take $\epsilon_{\text{2a}} = O(m^{3/2}n^{-1/2})$ and $\epsilon_{\text{2b-3}} = 0$. Lemma 16 now implies the first bound in the theorem statement. The second bound follows from an analogous argument for the distributions

$$\mathcal{P}_0 = \text{Bern}(q)^{\otimes m\times n}, \quad \mathcal{P}_1 = \mathcal{N}\left(0, I_m\right)^{\otimes n}, \quad \mathcal{P}_{\text{2a}} = \mathcal{W}_m\left(n, I_m\right) \quad \text{and} \quad \mathcal{P}_{\text{2b-3}} = \mathcal{N}\left(0, I_d\right)^{\otimes n}$$

with $\epsilon_1 = O(m^{-1}n^{-1})$, $\epsilon_{\text{2a}} = O(m^{3/2}n^{-1/2})$ and $\epsilon_{\text{2b-3}} = 0$. This completes the proof of the theorem. ■

H.2. Comparing Wishart and Inverse Wishart

This section is devoted to proving the upper bound on the KL divergence between Wishart matrices and their inverses in Theorem 44 used in the proof of Theorem 43. As noted in the previous subsection, the next theorem also implies total variation convergence between Wishart and inverse Wishart when $n \gg d^3$ by Pinsker's inequality. This theorem is related to a line of recent research examining the total variation convergence between ensembles of random matrices in the regime where $n \gg d$. A number of recent papers have investigated the total variation convergence between the fluctuations of the Wishart and Gaussian orthogonal ensembles, also showing these converge when $n \gg d^3$ (Jiang and Li, 2015; Bubeck et al., 2016; Bubeck and Ganguly, 2016; Rácz and Richey, 2019), convergence with other matrix ensembles at intermediate asymptotic scales of $d \ll n \ll d^3$ (Chételat and Wells, 2019) and applications of these results to random geometric graphs (Bubeck et al., 2016; Eldan and Mikulincer, 2016; Brennan et al., 2019b).

Let $\Gamma_d(x)$ and $\psi_d(x)$ denote the multivariate gamma and digamma functions given by

$$\Gamma_d(a) = \pi^{d(d-1)/4} \cdot \prod_{i=1}^{d} \Gamma\left(a - \frac{i-1}{2}\right) \quad \text{and} \quad \psi_d(a) = \frac{\partial \log \Gamma_d(a)}{\partial a} = \sum_{i=1}^{d} \psi\left(a - \frac{i-1}{2}\right)$$

where $\Gamma(z)$ and $\psi = \Gamma'(z)/\Gamma(z)$ denote the ordinary gamma and digamma functions. We will need several approximations to the log-gamma and digamma functions to prove our desired bound on KL

divergence. The classical Stirling series for the log-gamma function is

$$\log\Gamma(z) \sim \frac{1}{2}\log(2\pi) + \left(z - \frac{1}{2}\right)\log z - z + \sum_{k=1}^{\infty} \frac{B_{2k}}{2k(2k-1)z^{2k-1}}$$

where B_m denotes the mth Bernoulli number. While this series does not converge absolutely for any z because of the growth rate of the coefficients B_{2k}, its partial sums are increasingly accurate. More precisely, we have the following series approximation to the log-gamma function (see e.g. pg. 67 of Remmert (2013)) up to second order

$$\log\Gamma(z) = \frac{1}{2}\log(2\pi) + \left(z - \frac{1}{2}\right)\log z - z + \frac{1}{12z} + O(z^{-3})$$

as $z \to \infty$. A similar series expansion exists for the digamma function, given by

$$\psi(z) \sim \log z - \frac{1}{2z} - \sum_{k=1}^{\infty} \frac{B_{2k}}{2kz^{2k}}$$

This series also exhibits the phenomenon that, while not converge absolutely for any z, its partial sums are increasingly accurate. We have the following third order expansion of $\psi(z)$ given by

$$\psi(z) = \log z - \frac{1}{2z} - \frac{1}{12z^2} + \frac{2}{z^2}\int_0^{\infty} \frac{t^3}{(t^2+z^2)(e^{2\pi t}-1)}dt = \log z - \frac{1}{2z} - \frac{1}{12z^2} + O(z^{-4})$$

as $z \to \infty$. We now state and prove the main theorem of this section.

Theorem 44 (Comparing Wishart and Inverse Wishart) *Let $n \geq d+1$ and $m \geq d$ be positive integers such that $n = \Theta(m)$, $|m-n| = o(n)$ and $n - d = \Omega(n)$ as $m, n, d \to \infty$, and let $\beta = \frac{1}{m(n-d-1)}$. Then*

$$d_{KL}\left(\mathcal{W}_d(n, I_d) \,\middle\|\, \mathcal{W}_d^{-1}(m, \beta \cdot I_d)\right) = \frac{d^3}{6n} + \frac{s^2 d(d+1)}{8n^2} - \frac{5sd^3}{24n^2} + \frac{sd^3}{12mn} + O\left(d^2 n^{-3}|s|^3 + d^4 n^{-2} + d^2 n^{-1}\right)$$

where $s = n - m$. In particular, when $m = n$ and $n \gg d^3$ it follows that

$$d_{KL}\left(\mathcal{W}_d(n, I_d) \,\middle\|\, \mathcal{W}_d^{-1}(n, \beta \cdot I_d)\right) = o(1)$$

Proof Note that the given conditions also imply that $m - d = \Omega(m)$. Let $X \sim \mathcal{W}_d(n, I_d)$ and $Y \sim \mathcal{W}_d^{-1}(m, \beta \cdot I_d)$. Throughout this section, $A \in \mathbb{R}^{d \times d}$ will denote a positive semidefinite matrix. It is well known that the Wishart distribution $\mathcal{W}_d(n, I_d)$ is absolutely continuous with respect to the Lebesgue measure on the cone $\mathcal{C}_d^{\text{PSD}}$ of positive semidefinite matrices in $\mathbb{R}^{d \times d}$ (Wishart, 1928). Furthermore the density of X with respect to the Lebesgue measure can be written as

$$f_X(A) = \frac{1}{2^{nd/2} \cdot \Gamma_d\left(\frac{n}{2}\right)} \cdot |A|^{(n-d-1)/2} \cdot \exp\left(-\frac{1}{2}\text{Tr}(A)\right)$$

A change of variables from $A \to \beta^{-1} \cdot A^{-1}$ shows that the distribution $\mathcal{W}_d^{-1}(m, \beta \cdot I_d)$ is also absolutely continuous with respect to the Lebesgue measure on $\mathcal{C}_d^{\text{PSD}}$. It is well-known (see e.g. Gelman et al. (2013)) that the density of Y can be written as

$$f_Y(A) = \frac{\beta^{-md/2}}{2^{md/2} \cdot \Gamma_d\left(\frac{m}{2}\right)} \cdot |A|^{-(m+d+1)/2} \cdot \exp\left(-\frac{\beta^{-1}}{2} \cdot \text{Tr}\left(A^{-1}\right)\right)$$

Now note that

$$\begin{aligned}\log f_X(A) - \log f_Y(A) = \frac{(m-n)d}{2} \cdot \log 2 + \log \Gamma_d\left(\frac{m}{2}\right) - \log \Gamma_d\left(\frac{n}{2}\right) + \frac{md}{2} \cdot \log \beta \\ + \frac{m+n}{2} \cdot \log |A| - \frac{1}{2}\text{Tr}(A) + \frac{\beta^{-1}}{2} \cdot \text{Tr}\left(A^{-1}\right)\end{aligned}$$

The expectation of $\log |A|$ where $A \sim \mathcal{W}_d(n, I_d)$ is well known (e.g. see pg. 693 of Bishop (2006)) to be equal to

$$\mathbb{E}_{A \sim \mathcal{W}_d(n, I_d)}\left[\log |A|\right] = \psi_d\left(\frac{n}{2}\right) + d \log 2$$

Furthermore, it is well known (e.g. see pg. 85 of Mardia et al. (1979)) that the mean of A^{-1} if $A \sim \mathcal{W}_d(n, I_d)$ is

$$\mathbb{E}_{A \sim \mathcal{W}_d(n, I_d)}\left[A^{-1}\right] = \frac{I_d}{n-d-1}$$

Therefore we have that $\mathbb{E}_{A \sim \mathcal{W}_d(n, I_d)}\left[\text{Tr}\left(A^{-1}\right)\right] = d/(n-d-1)$. Similarly, we have that $\mathbb{E}_{A \sim \mathcal{W}_d(n, I_d)}\left[A\right] = n \cdot I_d$ and thus $\mathbb{E}_{A \sim \mathcal{W}_d(n, I_d)}\left[\text{Tr}(A)\right] = nd$. Combining these identities yields that

$$\begin{aligned}d_{\text{KL}}\left(\mathcal{W}_d(n, I_d) \,\middle\|\, \mathcal{W}_d^{-1}(m, \beta \cdot I_d)\right) &= \mathbb{E}_{A \sim \mathcal{W}_d(n, I_d)}\left[\log f_X(A) - \log f_Y(A)\right] \\ &= \frac{(m-n)d}{2} \cdot \log 2 + \log \Gamma_d\left(\frac{m}{2}\right) - \log \Gamma_d\left(\frac{n}{2}\right) + \frac{md}{2} \cdot \log \beta \\ &\quad + \frac{m+n}{2} \cdot \left(\psi_d\left(\frac{n}{2}\right) + d \log 2\right) - \frac{nd}{2} + \frac{\beta^{-1} d}{2(n-d-1)}\end{aligned} \tag{7}$$

We now use the series approximations for $\Gamma(z)$ and $\psi(z)$ mentioned above to approximate each of these terms. Note that since $m - d = \Omega(m)$, we have that

$$\begin{aligned}\log \Gamma_d\left(\frac{m}{2}\right) &= \frac{d(d-1)}{4} \log \pi + \sum_{i=1}^{d} \log \Gamma\left(\frac{m-i+1}{2}\right) \\ &= \frac{d(d-1)}{4} \log \pi + \sum_{i=1}^{d} \left(\frac{1}{2}\log(2\pi) + \left(\frac{m-i}{2}\right) \log\left(\frac{m-i+1}{2}\right)\right. \\ &\qquad \left. - \left(\frac{m-i+1}{2}\right) + \frac{1}{6(m-i+1)} + O(m^{-3})\right) \\ &= \frac{d(d-1)}{4} \log \pi + \frac{d}{2}\log(2\pi) - \frac{dm}{2} + \frac{d(d-1)}{4} + O(dm^{-3}) \\ &\quad + \sum_{i=1}^{d} \left(\left(\frac{m-i}{2}\right) \log\left(\frac{m}{2}\right) + \left(\frac{m-i}{2}\right) \log\left(1 - \frac{i-1}{m}\right) + \frac{1}{6(m-i+1)}\right)\end{aligned}$$

using the fact that $\sum_{i=1}^{d}(i-1) = d(d-1)/2$. Let H_n denote the harmonic series $H_n = \sum_{i=1}^{n} 1/i$. Using the well-known fact that $\psi(n+1) = H_n - \gamma$ where γ is the Euler-Mascheroni constant, we have that

$$\begin{aligned}\sum_{i=1}^{d} \frac{1}{m-i+1} &= H_m - H_{m-d} \\ &= \log(m+1) - \log(m-d+1) + O(m^{-1}) \\ &= \frac{d}{m+1} + \frac{d^2}{2(m+1)^2} + O(d^3 m^{-3}) + O(m^{-1}) \\ &= O\left(dm^{-1}\right)\end{aligned}$$

where the second last estimate follows applying the Taylor approximation $\log(1-x) = -x - \frac{1}{2}x^2 + O(x^3)$ for $x = \frac{d}{m+1} \in (0,1)$. Applying this Taylor approximation again, we have that

$$\begin{aligned}&\sum_{i=1}^{d} \left(\frac{m-i}{2}\right) \log\left(1 - \frac{i-1}{m}\right) \\ &\quad= -\frac{1}{2}\sum_{i=1}^{d} \left(\frac{(m-i)(i-1)}{m} + \frac{(m-i)(i-1)^2}{2m^2} + O\left(i^3 m^{-2}\right)\right) \\ &\quad= O(d^4 m^{-2}) - \frac{1}{2}\sum_{i=1}^{d} \left(\frac{(m-1)(i-1)}{m} - \frac{(i-1)^2}{m} + \frac{(m-1)(i-1)^2}{2m^2} - \frac{(i-1)^3}{2m^2}\right) \\ &\quad= O(d^4 m^{-2}) - \frac{(m-1)d(d-1)}{4m} + \frac{d(d-1)(2d-1)}{12m} - \frac{(m-1)d(d-1)(2d-1)}{24m^2} + \frac{d^2(d-1)^2}{16m^2} \\ &\quad= O(d^4 m^{-2}) - \frac{d(d-1)}{4} + \frac{d(d-1)(2d+5)}{24m}\end{aligned}$$

using the identities $\sum_{i=1}^{d}(i-1)^2 = d(d-1)(2d-1)/6$ and $\sum_{i=1}^{d}(i-1)^3 = d^2(d-1)^2/4$. Combining all of these approximations and simplifying using the fact that $m - d = \Omega(m)$ yields that

$$\begin{aligned}\log\Gamma_d\left(\frac{m}{2}\right) = \frac{d(d-1)}{4}\log\pi + \frac{d}{2}\log(2\pi) - \frac{dm}{2} + \frac{dm}{2}\log\left(\frac{m}{2}\right) - \frac{d(d+1)}{4}\log\left(\frac{m}{2}\right) \\ + \frac{d(d-1)(2d+5)}{24m} + O\left(d^4 m^{-2} + dm^{-1}\right)\end{aligned}$$

as $m, d \to \infty$ and $m - d = \Omega(m)$. An analogous estimate is also true for $\log\Gamma_d\left(\frac{n}{2}\right)$. Similar approximations now yield since $n - d = \Omega(n)$, we have that

$$\begin{aligned}\psi_d\left(\frac{n}{2}\right) &= \sum_{i=1}^{d}\left(\log\left(\frac{n-i+1}{2}\right) - \frac{1}{n-i+1} + O(n^{-2})\right) \\ &= d\log\left(\frac{n}{2}\right) + \sum_{i=1}^{d}\log\left(1 - \frac{i-1}{n}\right) - H_n + H_{n-d} + O(dn^{-2}) \\ &= d\log\left(\frac{n}{2}\right) - \sum_{i=1}^{d}\left(\frac{i-1}{n} + \frac{(i-1)^2}{2n^2} + O\left(i^3 n^{-3}\right)\right) - \frac{d}{n+1} - \frac{d^2}{2(n+1)^2}\end{aligned}$$

$$
\begin{aligned}
&\quad + O\left(d^3 n^{-3} + d n^{-2}\right) \\
&= d \log\left(\frac{n}{2}\right) - \frac{d(d-1)}{2n} - \frac{d(d-1)(2d-1)}{12n^2} - \frac{d}{n+1} + O\left(d^4 n^{-3} + d^2 n^{-2}\right)
\end{aligned}
$$

Here we have expanded $\psi(n+1) = H_n - \gamma$ to an additional order with the approximation

$$
\begin{aligned}
H_n - H_{n-d} &= \log(n+1) - \log(n-d+1) - \frac{1}{2(n+1)} + \frac{1}{2(n-d+1)} + O(n^{-2}) \\
&= \frac{d}{n+1} + \frac{d^2}{2(n+1)^2} + O(dn^{-2})
\end{aligned}
$$

Combining all of these estimates and simplifying with $\beta^{-1} = m(n-d-1)$ now yields that

$$
\begin{aligned}
&d_{\mathrm{KL}}\left(\mathcal{W}_d(n, I_d) \,\middle\|\, \mathcal{W}_d^{-1}(m, \beta \cdot I_d)\right) \\
&\quad = md \log 2 + \frac{md}{2} \log \beta - \frac{nd}{2} + \frac{\beta^{-1} d}{2(n-d-1)} + \log \Gamma_d\left(\frac{m}{2}\right) - \log \Gamma_d\left(\frac{n}{2}\right) + \frac{m+n}{2} \cdot \psi_d\left(\frac{n}{2}\right) \\
&\quad = md \log 2 + \frac{md}{2} \log \beta - \frac{nd}{2} + \frac{\beta^{-1} d}{2(n-d-1)} - \frac{d(m-n)}{2} + \frac{dm}{2} \log\left(\frac{m}{2}\right) - \frac{dn}{2} \log\left(\frac{n}{2}\right) \\
&\qquad - \frac{d(d+1)}{4} \log\left(\frac{m}{n}\right) + \frac{d(d-1)(2d+5)}{24} \cdot (m^{-1} - n^{-1}) + \frac{(m+n)d}{2} \log\left(\frac{n}{2}\right) \\
&\qquad - \frac{(m+n)d(d-1)}{4n} - \frac{(m+n)d(d-1)(2d-1)}{24n^2} - \frac{(m+n)d}{2(n+1)} + O\left(d^4 n^{-2} + d^2 n^{-1}\right) \\
&\quad = -\frac{(m+n)d(d-1)}{4n} - \frac{(m+n)d}{2(n+1)} - \frac{d(d+1)}{4} \log\left(\frac{m}{n}\right) - \frac{dm}{2} \log\left(1 - \frac{d+1}{n}\right) \\
&\qquad - \frac{(m+n)d(d-1)(2d-1)}{24n^2} + \frac{(n-m)d(d-1)(2d+5)}{24mn} + O\left(d^4 n^{-2} + d^2 n^{-1}\right) \\
&\quad = -\frac{(m+n)d(d+1)}{4n} + \frac{d(d+1)}{4}\left(\frac{n-m}{n} + \frac{(n-m)^2}{2n^2} + O\left(n^{-3}|s|^3\right)\right) \\
&\qquad + \frac{dm}{2}\left(\frac{d+1}{n} + \frac{(d+1)^2}{2n^2} + O(d^3 n^{-3})\right) - \frac{(m+n)d(d-1)(2d-1)}{24n^2} \\
&\qquad + \frac{sd(d-1)(2d+5)}{24mn} + O\left(d^4 n^{-2} + d^2 n^{-1}\right) \\
&\quad = \frac{d^3}{6n} + \frac{s^2 d(d+1)}{8n^2} - \frac{5sd^3}{24n^2} + \frac{sd^3}{12mn} + O\left(d^2 n^{-3}|s|^3 + d^4 n^{-2} + d^2 n^{-1}\right)
\end{aligned}
$$

In the fourth equality, we used the fact that $1/(n+1) = 1/n + O(n^{-2})$, that $s = n - m = o(n)$ and the Taylor approximation $\log(1-x) = -x - \frac{1}{2}x^2 + O(x^3)$ for $|x| < 1$. The last line follows from absorbing small terms into the error term. The second part of the theorem statement follows immediately from substituting $m = n$ and $s = 0$ into the bound above and noting that the dominant term is $d^3/6n$ when $n \gg d^3$. ∎

We now make two remarks on the theorem above. The first motivates the choice of the parameter β to satisfy $\beta^{-1} = m(n-d-1)$. Note that the KL divergence in Equation (7) depends on β through the terms

$$
\frac{md}{2} \log \beta + \frac{\beta^{-1} d}{2(n-d-1)}
$$

which is minimized at the stationary point $\beta^{-1} = m(n-d-1)$. Thus the KL divergence in Equation (7) is minimized for a fixed pair (m, n) at this value of β. We also remark that the distributions $\mathcal{W}_d(n, I_d)$ and $\mathcal{W}_d^{-1}(m, \beta \cdot I_d)$ only converge in KL divergence if $d \gg n^3$ as the expression in Theorem 44 is easily seen to not converge to zero if $d = O(n^3)$.

Appendix I. Negative Correlations, Sparse Mixtures and Supervised Problems

In the first part of this section, we introduce and give a reduction to the intermediate problem imbalanced sparse Gaussian mixtures, as outlined in Section C.3 and the beginning of Section G. This reduction is then used in the second part of this section, along with the reduction to negative sparse PCA in the previous section, as a subroutine in a reduction to robust sparse linear regression and mixtures of sparse linear regressions, as outlined in Section C.4. Our reduction to imbalanced sparse Gaussian mixtures will also be used in Section L to show computational lower bounds for robust sparse mean estimation.

I.1. Reduction to Imbalanced Sparse Gaussian Mixtures

In this section, we give our reduction from k-BPDS to the intermediate problem ISGM, which we will reduce from in subsequent sections to obtain several of our main computational lower bounds. We present our reduction to ISGM with dense Bernoulli rotations applied with the design matrix $K_{r,t}$ from Definition 28, and at the end of this section sketch the variant using the random design matrix alternative $R_{n,\epsilon}$ introduced in Section G.4. Throughout this section, the input k-BPDS instance will be described by its $m \times n$ adjacency matrix of Bernoulli random variables. The problem ISGM, imbalanced sparse Gaussian mixtures, is a simple vs. simple hypothesis testing problem defined formally below. A similar distribution was also used in Diakonikolas et al. (2017) to construct an instance of robust sparse mean estimation inducing the tight statistical-computational gap in the statistical query model.

Definition 45 (Imbalanced Sparse Gaussian Mixtures) *Given some $\mu \in \mathbb{R}$ and $\epsilon \in (0, 1)$, let μ' be such that $\epsilon \cdot \mu' + (1-\epsilon) \cdot \mu = 0$. For each subset $S \subseteq [d]$, $\text{ISGM}_D(n, S, d, \mu, \epsilon)$ denotes the distribution over $X = (X_1, X_2, \ldots, X_n)$ where $X_i \in \mathbb{R}^d$ where*

$$X_1, X_2, \ldots, X_n \sim_{\text{i.i.d.}} \text{MIX}_\epsilon \left(\mathcal{N}(\mu \cdot \mathbf{1}_S, I_d), \mathcal{N}(\mu' \cdot \mathbf{1}_S, I_d)\right)$$

We will use the notation $\text{ISGM}(n, k, d, \mu, \epsilon)$ to refer to the hypothesis testing problem between $H_0 : X_1, X_2, \ldots, X_n \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$ and an alternative hypothesis H_1 sampling the distribution above where S is chosen uniformly at random among all k-subsets of $[d]$. Our reduction k-BPDS-TO-ISGM is shown in Figure 9. The next theorem encapsulates the total variation guarantees of this reduction. A key parameter is the prime number r, which is used to parameterize the design matrices $K_{r,t}$ in the BERN-ROTATIONS step.

To show the tightest possible statistical-computational gaps in applications of this theorem, we ideally would want to take n such that $n = \Theta(k_n r^t)$. When r is growing with N, this induces number theoretic constraints on our choices of parameters that require careful attention and will be discussed in Section L.1. Because of this subtlety, we have kept the statement of our next theorem technically precise and in terms of all of the free parameters of the reduction k-BPDS-TO-ISGM. Ignoring these number theoretic constraints, the reduction k-BPDS-TO-ISGM can be interpreted as

Algorithm k-BPDS-TO-ISGM

Inputs: Matrix $M \in \{0,1\}^{m \times n}$, dense subgraph dimensions k_m and k_n where k_n divides n and the following parameters

- partition F of $[n]$ into k_n parts of size n/k_n, edge probabilities $0 < q < p \leq 1$ and a slow growing function $w(n) = \omega(1)$
- target ISGM parameters (N, d, μ, ϵ) satisfying that $\epsilon = 1/r$ for some prime number r,

$$wN \leq k_n r \ell, \quad m \leq d, \quad n \leq k_n r^t \leq \text{poly}(n) \quad \text{and} \quad \mu \leq \frac{c}{\sqrt{r^t(r-1)\log(k_n m r^t)}}$$

for some $t \in \mathbb{N}$, a sufficiently small constant $c > 0$ and where $\ell = \frac{r^t - 1}{r-1}$

1. *Pad*: Form $M_{\text{PD}} \in \{0,1\}^{m \times k_n r^t}$ by adding $k_n r^t - n$ new columns sampled i.i.d. from $\text{Bern}(q)^{\otimes m}$ to the right end of M. Let F' be the partition formed by letting F'_i be F_i with exactly $r^t - n/k_n$ of the new columns.
2. *Bernoulli Rotations*: Fix a partition $[k_n r\ell] = F''_1 \cup F''_2 \cup \cdots \cup F''_{k_n}$ into k_n parts each of size $r\ell$ and compute the matrix $M_{\text{R}} \in \mathbb{R}^{m \times k_n r\ell}$ as follows:
   (1) For each row i and part F'_j, apply BERN-ROTATIONS to the vector $(M_{\text{PD}})_{i,F'_j}$ of entries in row i and in columns from F'_j with matrix parameter $K_{r,t}$, rejection kernel parameter $R_{\text{RK}} = k_n m r^t$, Bernoulli probabilities $0 < q < p \leq 1$, $\lambda = \sqrt{1 + (r-1)^{-1}}$, mean parameter $\lambda\sqrt{r^t(r-1)} \cdot \mu$ and output dimension $r\ell$.
   (2) Set the entries of $(M_{\text{R}})_{i,F''_j}$ to be the entries in order of the vector output in (1).
3. *Permute and Output*: Form $X \in \mathbb{R}^{d \times N}$ by choosing N distinct columns of M_{R} uniformly at random, embedding the resulting matrix as the first m rows of X and sampling the remaining $d - m$ rows of X i.i.d. from $\mathcal{N}(0, I_N)$. Output the columns $(X_1, X_2, \ldots, X_N)$ of X.

Figure 9: Reduction from bipartite k-partite planted dense subgraph to exactly imbalanced sparse Gaussian mixtures.

essentially mapping an instance of k-BPDS with parameters (m, n, k_m, k_n, p, q) with $k_n = o(\sqrt{n})$, $k_m = o(\sqrt{m})$ and planted row indices S where $|S| = k_m$ to the instance $\text{ISGM}_D(N, S, d, \mu, \epsilon)$ where $\epsilon \in (0,1)$ is arbitrary and can vary with n. The target parameters N, d and μ satisfy that

$$d = \Omega(m), \quad N = o(n) \quad \text{and} \quad \mu \asymp \frac{1}{\sqrt{\log n}} \cdot \sqrt{\frac{\epsilon k_n}{n}}$$

All of our applications will handle the number theoretic constraints to set parameters so that they nearly satisfy these conditions. The slow-growing function $w(n)$ is so that Step 3 subsamples the produced samples by a large enough factor to enable an application of finite de Finetti's theorem.

We now state our total variation guarantees for k-BPDS-TO-ISGM. Given a partition F of $[n]$ with $[n] = F_1 \cup F_2 \cup \cdots \cup F_{k_n}$, let $\mathcal{U}_n(F)$ denote the distribution of k_n-subsets of $[n]$ formed by choosing one member element of each of $F_1, F_2, \ldots, F_{k_n}$ uniformly at random. Let $\mathcal{U}_{n,k_n}$ denote the uniform distribution on k_n-subsets of $[n]$.

Theorem 46 (Reduction from k-BPDS to ISGM) *Let n be a parameter, $r = r(n) \geq 2$ be a prime number and $w(n) = \omega(1)$ be a slow-growing function. Fix initial and target parameters as follows:*

- Initial k-BPDS Parameters: *vertex counts on each side m and n that are polynomial in one another, dense subgraph dimensions k_m and k_n where k_n divides n, edge probabilities $0 < q < p \leq 1$ with $\min\{q, 1-q\} = \Omega(1)$ and $p - q \geq (mn)^{-O(1)}$, and a partition F of $[n]$.*
- Target ISGM Parameters: (N, d, μ, ϵ) *where $\epsilon = 1/r$ and there is a parameter $t = t(N) \in \mathbb{N}$ with*

$$wN \leq \frac{k_n r(r^t - 1)}{r - 1}, \quad m \leq d \leq \text{poly}(n), \quad n \leq k_n r^t \leq \text{poly}(n) \quad \text{and}$$

$$0 \leq \mu \leq \frac{\delta}{2\sqrt{6\log(k_n m r^t) + 2\log(p-q)^{-1}}} \cdot \frac{1}{\sqrt{r^t(r-1)(1+(r-1)^{-1})}}$$

where $\delta = \min\left\{\log\left(\frac{p}{q}\right), \log\left(\frac{1-q}{1-p}\right)\right\}$.

Let $\mathcal{A}(G)$ denote k-BPDS-TO-ISGM applied with the parameters above to a bipartite graph G with m left vertices and n right vertices. Then $\mathcal{A}$ runs in $\text{poly}(m, n)$ time and it follows that

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S \times T, p, q)\right), \text{ISGM}_D(N, S, d, \mu, \epsilon)\right) = O\left(w^{-1} + k_n^{-2} m^{-2} r^{-2t}\right)$$
$$d_{TV}\left(\mathcal{A}\left(\text{Bern}(q)^{\otimes m \times n}\right), \mathcal{N}(0, I_d)^{\otimes N}\right) = O\left(k_n^{-2} m^{-2} r^{-2t}\right)$$

for all subsets $S \subseteq [m]$ with $|S| = k_m$ and subsets $T \subseteq [n]$ with $|T| = k_n$ and $|T \cap F_i| = 1$ for each $1 \leq i \leq k_n$.

In the rest of this section, let $\mathcal{A}$ denote the reduction k-BPDS-TO-ISGM with input (M, F) where F is a partition of $[n]$ and output $(X_1, X_2, \ldots, X_N)$. Let $\text{Hyp}(N, K, n)$ denote a hypergeometric distribution with n draws from a population of size N with K success states. We will also need the upper bound on the total variation between hypergeometric and binomial distributions given by

$$d_{\text{TV}}\left(\text{Hyp}(N, K, n), \text{Bin}(n, K/N)\right) \leq \frac{4n}{N}$$

This bound is a simple case of finite de Finetti's theorem and is proven in Theorem (4) in Diaconis and Freedman (1980). We now proceed to establish the total variation guarantees for Bernoulli rotations and subsampling as in Steps 2 and 3 of $\mathcal{A}$ in the next two lemmas.

Before proceeding to prove these lemmas, we make a definition that will be used in the next few sections. Suppose that M is a $b \times a$ matrix, F and F' are partitions of $[ka]$ and $[kb]$ into k equally sized parts and $S \subseteq [kb]$ is such that $|S \cap F_i| = 1$ for each $1 \leq i \leq k$. Then define the vector $v = v_{S,F,F'}(M) \in \mathbb{R}^{kb}$ to be such that the restriction $v_{F_i'}$ to the elements of F_i' is given by

$$v_{F_i'} = M_{\cdot, \sigma_{F_i}(j)} \quad \text{where } j \text{ is the unique element in } S \cap F_i$$

Here, $M_{\cdot,j}$ denotes the jth column of M and σ_{F_i} denotes the order preserving bijection from F_i to $[b]$. In other words, $v_{S,F,F'}$ is the vector formed by concatenating the columns of M along the partition F', where the elements $S \cap F_i$ select which column appears along each part F'_i. In this section, whenever $S \cap F_i$ has size one, we will abuse notation and also use $S \cap F_i$ to denote its unique element.

Lemma 47 (Bernoulli Rotations for ISGM) *Let F' and F'' be a fixed partitions of $[k_n r^t]$ and $[k_n r\ell]$ into k_n parts of size r^t and $r\ell$, respectively, and let $S \subseteq [m]$ be a fixed k_m-subset. Let $T \subseteq [k_n r^t]$ where $|T \cap F'_i| = 1$ for each $1 \le i \le k_n$. Let $\mathcal{A}_2$ denote Step 2 of k-BPDS-TO-ISGM with input M_{PD} and output M_{R}. Suppose that p, q and μ are as in Theorem 46, then it follows that*

$$d_{TV}\left(\mathcal{A}_2\left(\mathcal{M}_{[m]\times[k_n r^t]}\left(S \times T, \text{Bern}(p), \text{Bern}(q)\right)\right),\right.$$
$$\left.\mathcal{L}\left(\mu\sqrt{r^t(r-1)} \cdot \mathbf{1}_S v_{T,F',F''}(K_{r,t})^\top + \mathcal{N}(0,1)^{\otimes m \times k_n r\ell}\right)\right) = O\left(k_n^{-2} m^{-2} r^{-2t}\right)$$
$$d_{TV}\left(\mathcal{A}_2\left(\text{Bern}(q)^{\otimes m \times k_n r^t}\right), \mathcal{N}(0,1)^{\otimes m \times k_n r\ell}\right) = O\left(k_n^{-2} m^{-2} r^{-2t}\right)$$

Proof First consider the case where $M_{\text{PD}} \sim \mathcal{M}_{[m]\times[k_n r^t]}\left(S \times T, \text{Bern}(p), \text{Bern}(q)\right)$. Observe that the subvectors of M_{PD} are distributed as

$$(M_{\text{PD}})_{i,F'_j} \sim \begin{cases} \text{PB}\left(F'_j, T \cap F'_j, p, q\right) & \text{if } i \in S \\ \text{Bern}(q)^{\otimes r^t} & \text{otherwise} \end{cases}$$

and are independent. Combining upper bound on the singular values of $K_{r,t}$ in Lemma 30, Lemma 26 applied with $R_{\text{RK}} = k_n m r^t$ and the condition on μ in the statement of Theorem 46 implies that

$$d_{\text{TV}}\left((M_{\text{R}})_{i,F''_j}, \mathcal{N}\left(\mu\sqrt{r^t(r-1)} \cdot (K_{r,t})_{\cdot,T\cap F'_j}, I_{r\ell}\right)\right) = O\left(k_n^{-3} m^{-3} r^{-2t}\right) \quad \text{if } i \in S$$
$$d_{\text{TV}}\left((M_{\text{R}})_{i,F''_j}, \mathcal{N}\left(0, I_{r\ell}\right)\right) = O\left(k_n^{-3} m^{-3} r^{-2t}\right) \quad \text{otherwise}$$

Now observe that the subvectors $(M_{\text{R}})_{i,F''_j}$ are also independent. Therefore the tensorization property of total variation in Fact 15 implies that $d_{\text{TV}}\left(M_{\text{R}}, \mathcal{L}(Z)\right) = O\left(k_n^{-2} m^{-2} r^{-2t}\right)$ where Z is defined so that its subvectors Z_{i,F''_j} are independent and distributed as

$$Z_{i,F''_j} \sim \begin{cases} \mathcal{N}\left(\mu\sqrt{r^t(r-1)} \cdot (K_{r,t})_{\cdot,T\cap F'_j}, I_{r\ell}\right) & \text{if } i \in S \\ \mathcal{N}\left(0, I_{r\ell}\right) & \text{otherwise} \end{cases}$$

Note that the entries of Z are independent Gaussians each with variance 1. Furthermore, the mean of Z can be verified to be exactly $\mu\sqrt{r^t(r-1)} \cdot \mathbf{1}_S v_{T,F',F''}(K_{r,t})^\top$. This completes the proof of the first total variation upper bound in the statement of the lemma. The second bound follows from the same argument above applied with $S = \emptyset$. ■

Lemma 48 (Subsampling for ISGM) *Let F', F'', S and T be as in Lemma 47. Let $\mathcal{A}_3$ denote Step 3 of k-PDS-TO-ISGM with input M_{R} and output $(X_1, X_2, \ldots, X_N)$. Then*

$$d_{TV}\left(\mathcal{A}_3\left(\tau \cdot \mathbf{1}_S v_{T,F',F''}(K_{r,t})^\top + \mathcal{N}(0,1)^{\otimes m \times k_n r\ell}\right), \text{ISGM}_D(N, S, d, \mu, \epsilon)\right) \le 4w^{-1}$$

where $\epsilon = 1/r$ and $\mu = \frac{\tau}{\sqrt{r^t(r-1)}}$. Furthermore, it holds that $\mathcal{A}_3\left(\mathcal{N}(0,1)^{\otimes m \times k_n r\ell}\right) \sim \mathcal{N}(0, I_d)^{\otimes N}$.

Proof Suppose that $M_{\mathrm{R}} \sim \tau \cdot \mathbf{1}_S K_{T,F',F''}^{\top} + \mathcal{N}(0,1)^{\otimes m \times k_n r\ell}$. For fixed S, T, F' and F'', the entries of M_{R} are independent. Observe that the columns of M_{R} are independent and either distributed according $\mathcal{N}(\mu \cdot \mathbf{1}_S, I_m)$ or $\mathcal{N}(\mu' \cdot \mathbf{1}_S, I_m)$ where $\mu' = \tau(1-r)/\sqrt{r^t(r-1)}$ depending on whether the entry of $v_{T,F',F''}(K_{r,t})$ at the index corresponding to the column is $1/\sqrt{r^t(r-1)}$ or $(1-r)/\sqrt{r^t(r-1)}$.

By Lemma 29, it follows that each column of $K_{r,t}$ contains exactly ℓ entries equal to $(1-r)/\sqrt{r^t(r-1)}$. This implies that exactly $k_n(r-1)\ell$ entries of $v_{T,F',F''}(K_{r,t})$ are equal to the value $1/\sqrt{r^t(r-1)}$. Define $\mathcal{R}_N(s)$ to be the distribution on $\mathbb{R}^N$ with a sample $v \sim \mathcal{R}_N(s)$ generated by first choosing an s-subset U of $[N]$ uniformly at random and then setting $v_i = 1/\sqrt{r^t(r-1)}$ if $i \in U$ and $v_i = (1-r)/\sqrt{r^t(r-1)}$ if $i \notin U$. Note that the number of columns distributed as $\mathcal{N}(\mu \cdot \mathbf{1}_S, I_m)$ in M_{R} chosen to be in X is distributed according to $\mathrm{Hyp}(k_n r\ell, k_n(r-1)\ell, N)$. Step 3 of $\mathcal{A}$ therefore ensures that, if M_{R} is distributed as above, then

$$X \sim \mathcal{L}\left(\tau \cdot \mathbf{1}_S \mathcal{R}_N(\mathrm{Hyp}(k_n\ell, k_n(r-1)\ell, N))^{\top} + \mathcal{N}(0,1)^{\otimes d \times N}\right)$$

Observe that the data matrix for a sample from $\mathrm{ISGM}_D(N, S, d, \mu, \epsilon)$ can be expressed similarly as

$$\mathrm{ISGM}_D(N, S, d, \mu, \epsilon) = \mathcal{L}\left(\tau \cdot \mathbf{1}_S \mathcal{R}_n(\mathrm{Bin}(N, 1-\epsilon))^{\top} + \mathcal{N}(0,1)^{\otimes d \times N}\right)$$

where again we set $\mu = \tau/\sqrt{r^t(r-1)}$. The conditioning property of d_{TV} in Fact 15 now implies that

$$d_{\mathrm{TV}}\left(\mathcal{L}(X), \mathrm{ISGM}_D(N, S, d, \mu, \epsilon)\right) \leq d_{\mathrm{TV}}\left(\mathrm{Bin}(N, 1-\epsilon), \mathrm{Hyp}\left(k_n r\ell, k_n(r-1)\ell, N\right)\right) \leq \frac{4N}{k_n r\ell} \leq 4w^{-1}$$

The last inequality follows from the application of Theorem (4) in Diaconis and Freedman (1980) to hypergeometric distributions above along with the fact that $1 - \epsilon = (k_n(r-1)\ell)/k_n r\ell$ and $wN \leq k_n r\ell$. This completes the proof of the upper bound in the lemma statement. Now consider applying the above argument with $\tau = 0$. It follows that $\mathcal{A}_3\left(\mathcal{N}(0,1)^{\otimes m \times k_n r\ell}\right) \sim \mathcal{N}(0,1)^{\otimes d \times N} = \mathcal{N}(0, I_d)^{\otimes N}$, which completes the proof of the lemma. ■

We now combine these lemmas to complete the proof of Theorem 46.

Proof [Proof of Theorem 46] We apply Lemma 16 to the steps $\mathcal{A}_i$ of $\mathcal{A}$ under each of H_0 and H_1 to prove Theorem 46. Define the steps of $\mathcal{A}$ to map inputs to outputs as follows

$$(M, F) \xrightarrow{\mathcal{A}_1} (M_{\mathrm{PD}}, F') \xrightarrow{\mathcal{A}_2} (M_{\mathrm{R}}, F'') \xrightarrow{\mathcal{A}_3} (X_1, X_2, \ldots, X_N)$$

We first prove the desired result in the case that H_1 holds. Consider Lemma 16 applied to the steps $\mathcal{A}_i$ above and the following sequence of distributions

$$\begin{aligned}
\mathcal{P}_0 &= \mathcal{M}_{[m]\times[n]}(S \times T, \mathrm{Bern}(p), \mathrm{Bern}(q)) \\
\mathcal{P}_1 &= \mathcal{M}_{[m]\times[k_n r^t]}(S \times T, \mathrm{Bern}(p), \mathrm{Bern}(q)) \\
\mathcal{P}_2 &= \mu\sqrt{r^t(r-1)} \cdot \mathbf{1}_S v_{T,F',F''}(K_{r,t})^{\top} + \mathcal{N}(0,1)^{\otimes m \times k_n r\ell} \\
\mathcal{P}_3 &= \mathrm{ISGM}_D(N, S, d, \mu, \epsilon)
\end{aligned}$$

As in the statement of Lemma 16, let ϵ_i be any real numbers satisfying $d_{\mathrm{TV}}\left(\mathcal{A}_i(\mathcal{P}_{i-1}), \mathcal{P}_i\right) \leq \epsilon_i$ for each step i. By construction, the step $\mathcal{A}_1$ is exact and we can take $\epsilon_1 = 0$. Lemma 47 yields that

we can take $\epsilon_2 = O\left(k_n^{-2} m^{-2} r^{-2t}\right)$. Applying Lemma 48 yields that we can take $\epsilon_3 = 4w^{-1}$. By Lemma 16, we therefore have that

$$d_{\mathrm{TV}}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S \times T, p, q)\right), \mathrm{ISGM}_D(N, S, d, \mu, \epsilon)\right) = O\left(w^{-1} + k_n^{-2} m^{-2} r^{-2t}\right)$$

which proves the desired result in the case of H_1. Now consider the case that H_0 holds and Lemma 16 applied to the steps $\mathcal{A}_i$ and the following sequence of distributions

$$\mathcal{P}_0 = \mathrm{Bern}(Q)^{\otimes m \times n}, \quad \mathcal{P}_1 = \mathrm{Bern}(Q)^{\otimes m \times k_n r^t}, \quad \mathcal{P}_2 = \mathcal{N}(0,1)^{\otimes m \times k_n r \ell} \quad \text{and} \quad \mathcal{P}_3 = \mathcal{N}(0, I_d)^{\otimes N}$$

As above, Lemmas 47 and 48 imply that we can take

$$\epsilon_1 = 0, \quad \epsilon_2 = O\left(k_n^{-2} m^{-2} r^{-2t}\right) \quad \text{and} \quad \epsilon_3 = 0$$

By Lemma 16, we therefore have that

$$d_{\mathrm{TV}}\left(\mathcal{A}\left(\mathrm{Bern}(q)^{\otimes m \times n}\right), \mathcal{N}(0, I_d)^{\otimes N}\right) = O\left(k_n^{-2} m^{-2} r^{-2t}\right)$$

which completes the proof of the theorem. ■

As discussed in Section G.4, we can replace $K_{r,t}$ in k-BPDS-TO-ISGM with the random matrix alternative $R_{L,\epsilon}$. More precisely, let k-BPDS-TO-ISGM$_R$ denote the reduction in Figure 9 with the following changes:

- At the beginning of the reduction, rejection sample $R_{L,\epsilon}$ for at most $\Theta((\log L)^2)$ iterations until the criteria of Lemma 38 are met, as outlined in Section G.4. Let $A \in \mathbb{R}^{L \times L}$ be the resulting matrix or stop the reduction if no such matrix is found. The latter case contributes $L^{-\omega(1)}$ to each of the total variation errors in Corollary 49.
- The dimensions $r\ell$ and r^t of the matrix $K_{r,t}$ used in BERN-ROTATIONS in Step 2 are both replaced throughout the reduction by the parameter L. This changes the output dimensions of M_{PD} and M_{R} in Steps 1 and 2 to both be $m \times k_n L$.
- In Step 2, apply BERN-ROTATIONS with A instead of $K_{r,t}$ and let $\lambda = C$ where C is the constant in Lemma 38.

The reduction k-BPDS-TO-ISGM$_R$ eliminates a number-theoretic constraint in k-BPDS-TO-ISGM arising from the fact the intermediate matrix M_{R} has a dimension that must be of the form $k_n r^t$ for some integer t. In contrast, k-BPDS-TO-ISGM$_R$ only requires that this dimension of M_{R} be a multiple of k_n. This will remove the condition (T) from our computational lower bounds for RSME, which is only restrictive in the very small ϵ regime of $\epsilon = n^{-\Omega(1)}$. We will deduce this computational lower bound for RSME implied by the reduction k-BPDS-TO-ISGM$_R$ formally in Section L.1.

The reduction k-BPDS-TO-ISGM$_R$ can be analyzed using an argument identical to the one above, with Lemma 38 used in place of Lemma 30 and accounting for the additional $L^{-\omega(1)}$ total variation error incurred by failing to obtain a $R_{n,\epsilon}$ satisfying the criteria in Lemma 38. Carrying this out yields the following corollary. We remark that the new condition $\epsilon \gg L^{-1} \log L$ in the corollary below will amount to the condition $\epsilon \gg N^{-1/2} \log N$ in our computational lower bounds. This is because, in our applications, we will typically set $N = \tilde{\Theta}(k_n L)$ and k_n to be very close to but slightly smaller than $\sqrt{n} = \tilde{\Theta}(\sqrt{N})$, to ensure that the input k-BPDS instance is hard. These conditions together with $\epsilon \gg L^{-1} \log L$ amount to the condition on the target parameters given by $\epsilon \gg N^{-1/2} \log N$.

Corollary 49 (Reduction from k-BPDS to ISGM with $R_{L,\epsilon}$) *Let n be a parameter and let $w(n) = \omega(1)$ be a slow-growing function. Fix initial and target parameters as follows:*

- Initial k-BPDS Parameters: m, n, k_m, k_n, p, q *and* F *as in Theorem 46.*
- Target ISGM Parameters: (N, d, μ, ϵ) *such that there is a parameter* $L = L(N) \in \mathbb{N}$ *such that* $L(N) \to \infty$ *and it holds that*

$$\max\{wN, n\} \le k_n L \le \text{poly}(n), \quad m \le d \le \text{poly}(n), \quad \frac{w \log L}{L} \le \epsilon \le \frac{1}{2} \quad \text{and}$$

$$0 \le \mu \le \frac{C\delta}{\sqrt{\log(k_n m L) + \log(p-q)^{-1}}} \cdot \sqrt{\frac{\epsilon}{L}}$$

for some sufficiently small constant $C > 0$, where δ is as in Theorem 46.

If $\mathcal{A}$ denotes k-BPDS-TO-ISGM$_R$ applied with the parameters above, then $\mathcal{A}$ runs in poly(m, n) *time and*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S \times T, p, q)\right), \text{ISGM}_D(N, S, d, \mu, \epsilon)\right) = o(1)$$
$$d_{TV}\left(\mathcal{A}\left(\text{Bern}(q)^{\otimes m \times n}\right), \mathcal{N}(0, I_d)^{\otimes N}\right) = o(1)$$

for all k_m-subsets $S \subseteq [m]$ and k_n-subsets $T \subseteq [n]$ with $|T \cap F_i| = 1$ for each $1 \le i \le k_n$.

I.2. Sparse Mixtures of Regressions and Negative Sparse PCA

In this section, we combine the previous two reductions to NEG-SPCA and ISGM with some additional observations to produce a single reduction that will be used to prove two of our main results in Section L.3 – computational lower bounds for mixtures of SLRs and robust SLR. We begin this section by generalizing our definition of the distribution $\text{MSLR}_D(n, S, d, \gamma, 1/2)$ from Section E.3 to simultaneously capture the mixtures of SLRs distributions we will reduce to and our adversarial construction for robust SLR.

Recall from Section E.3 that $\text{LR}_d(v)$ denotes the distribution of a single sample-label pair $(X, y) \in \mathbb{R}^d \times \mathbb{R}$ given by $y = \langle v, X \rangle + \eta$ where $X \sim \mathcal{N}(0, I_d)$ and $\eta \sim \mathcal{N}(0, 1)$. Our generalization of MSLR_D will be parameterized by $\epsilon \in (0, 1)$. The canonical setup for mixtures of SLRs from Section E.3 corresponds to setting $\epsilon = 1/2$ and formally is restated in the following definition for convenience.

Definition 50 (Mixtures of Sparse Linear Regressions with $\epsilon = 1/2$) *Let $\gamma \in \mathbb{R}$ be such that $\gamma > 0$. For each subset $S \subseteq [d]$, let* $\text{MSLR}_D(n, S, d, \gamma, 1/2)$ *denote the distribution over n-tuples of independent data-label pairs $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ where $X_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ are sampled as follows:*

- *first sample n independent Rademacher random variables $s_1, s_2, \ldots, s_n \sim_{\text{i.i.d.}}$* Rad*; and*
- *then form data-label pairs $(X_i, y_i) \sim \text{LR}_d(\gamma s_i v_S)$ for each $1 \le i \le n$.*

where $v_S \in \mathbb{R}^d$ is the $|S|$-sparse unit vector $v_S = |S|^{-1/2} \cdot \mathbf{1}_S$.

Algorithm k-BPDS-TO-MSLR

Inputs: Matrix $M \in \{0,1\}^{m \times n}$, dense subgraph dimensions k_m and k_n where k_n divides n and the following parameters

- partition F, edge probabilities $0 < q < p \le 1$ and $w(n)$ as in Figure 9
- target MSLR parameters (N, d, γ, ϵ) and prime r and $t \in \mathbb{N}$ where N, d, r, t, ℓ and $\epsilon = 1/r$ are as in Figure 9 with the additional requirement that $N \le n$ and where $\gamma \in (0,1)$ satisfies that
$$\gamma^2 \le c \cdot \min\left\{ \frac{k_m}{r^{t+1}\log(k_n m r^t)\log N}, \frac{k_n k_m}{n \log(mn)} \right\}$$
for a sufficiently small constant $c > 0$.

1. *Clone*: Compute the matrices $M_{\text{ISGM}} \in \{0,1\}^{m \times n}$ and $M_{\text{NEG-SPCA}} \in \{0,1\}^{m \times n}$ by applying BERNOULLI-CLONE with $t = 2$ copies to the entries of the matrix M with input Bernoulli probabilities p and q, and output probabilities p and $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}}\left(\sqrt{q} - 1\right)$.
2. *Produce* ISGM *Instance*: Form $(Z_1, Z_2, \ldots, Z_N)$ where $Z_i \in \mathbb{R}^d$ as the output of k-BPDS-TO-ISGM applied to the matrix M_{ISGM} with partition F, edge probabilities $0 < Q < p \le 1$, slow-growing function w, target ISGM parameters (N, d, μ, ϵ) and $\mu > 0$ given by $\mu = 4\gamma \cdot \sqrt{\frac{\log N}{k_m}}$.
3. *Produce* NEG-SPCA *Instance*: Form $(W_1, W_2, \ldots, W_n)$ where $W_i \in \mathbb{R}^d$ as the output of BPDS-TO-NEG-SPCA applied to the matrix $M_{\text{NEG-SPCA}}$ with edge probabilities $0 < Q < p \le 1$, target dimension d and parameter $\tau > 0$ satisfying that $\tau^2 = \frac{8n\gamma^2}{k_n k_m (1-\gamma^2)}$.
4. *Scale and Label* ISGM *Instance*: Generate $y_1, y_2, \ldots, y_N \sim_{\text{i.i.d.}} \mathcal{N}(0, 1+\gamma^2)$ and truncate each y_i to satisfy $|y_i| \le 2\sqrt{(1+\gamma^2)\log N}$. Generate $G_1, G_2, \ldots, G_N \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$ and form $(Z'_1, Z'_2, \ldots, Z'_N)$ where $Z'_i \in \mathbb{R}^d$ as
$$Z'_i = \frac{y_i}{4(1+\gamma^2)}\sqrt{\frac{2}{\log N}} \cdot Z_i + \sqrt{1 - \frac{y_i^2}{4(1+\gamma^2)^2 \log N}} \cdot G_i$$
5. *Merge and Output*: For each $1 \le i \le N$, let $X_i = \frac{1}{\sqrt{2}}\left(Z'_i + W_i\right)$ and output the N labelled pairs $(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)$.

Figure 10: Reduction from bipartite planted dense subgraph to mixtures of sparse linear regressions through imbalanced Gaussian mixtures and negative sparse PCA

Our more general formulation when $\epsilon < 1/2$ is described in the definition below. When $\epsilon < 1/2$, the distribution $\text{MSLR}_D(n, S, d, \gamma, \epsilon)$ can always be produced by an adversary in robust SLR. This observation will be discussed in more detail and used in Section L.3 to show com-

putational lower bounds for robust SLR. The reason we have chosen to write these two different distributions under a common notation is that the main reduction of this section, k-BPDS-TO-MSLR, will simultaneously map to both mixtures of SLRs and robust SLR. Lower bounds for the mixture problem will be obtained by setting $r = 2$ in the reduction to ISGM used as a subroutine in k-BPDS-TO-MSLR, while lower bounds for robust sparse regression will be obtained by taking $r > 2$. These implications of k-BPDS-TO-MSLR are discussed further in Section L.

Definition 51 (Mixtures of Sparse Linear Regressions with $\epsilon < 1/2$) *Let $\gamma > 0$, $\epsilon \in (0, 1/2)$ and let a denote $a = \epsilon^{-1}(1-\epsilon)$. For each subset $S \subseteq [d]$, let $\mathrm{MSLR}_D(n, S, d, \gamma, \epsilon)$ denote the distribution over n-tuples of data-label pairs $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ sampled as follows:*

- *the pairs $(b_1, X_1, y_1), (b_2, X_2, y_2), \ldots, (b_n, X_n, y_n)$ are i.i.d. and $b_1, b_2, \ldots, b_n \sim_{\text{i.i.d.}} \mathrm{Bern}(1-\epsilon)$;*
- *if $b_i = 1$, then $(X_i, y_u) \sim \mathrm{LR}_d(\gamma v_S)$ where v_S is as in Definition 50; and*
- *if $b_i = 0$, then (X_i, y_i) is jointly Gaussian with mean zero and $(d+1) \times (d+1)$ covariance matrix*

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{Xy} \\ \Sigma_{yX} & \Sigma_{yy} \end{bmatrix} = \begin{bmatrix} I_d + \frac{(a^2-1)\gamma^2}{1+\gamma^2} \cdot v_S v_S^\top & -a\gamma \cdot v_S \\ -a\gamma \cdot v_S^\top & 1+\gamma^2 \end{bmatrix}$$

The main reduction of this section from k-BPDS to MSLR is shown in Figure 10. This reduction inherits the number theoretic constraints of our reduction to ISGM mentioned in the previous section. These will be discussed in more detail when k-BPDS-TO-MSLR is used to deduce computational lower bounds in Section L.3. The following theorem gives the total variation guarantees for k-BPDS-TO-MSLR.

Theorem 52 (Reduction from k-BPDS to MSLR) *Let n be a parameter, $r = r(n) \geq 2$ be a prime number and $w(n) = \omega(1)$ be a slow-growing function. Fix initial and target parameters as follows:*

- Initial k-BPDS Parameters: *vertex counts on each side m and n that are polynomial in one another and satisfy the condition that $n \gg m^3$, subgraph dimensions k_m and k_n where k_n divides n, constant densities $0 < q < p \leq 1$ and a partition F of $[n]$.*
- Target MSLR Parameters: (N, d, γ, ϵ) *where $\epsilon = 1/r$ and there is a parameter $t = t(N) \in \mathbb{N}$ with*

$$N \leq n, \quad wN \leq \frac{k_n r(r^t - 1)}{r-1}, \quad m \leq d \leq \mathrm{poly}(n), \quad \text{and} \quad n \leq k_n r^t \leq \mathrm{poly}(n)$$

and where $\gamma \in (0, 1/2)$ satisfies that

$$\gamma^2 \leq c \cdot \min\left\{ \frac{k_m}{r^{t+1} \log(k_n m r^t) \log N}, \frac{k_n k_m}{n \log(mn)} \right\}$$

for a sufficiently small constant $c > 0$.

Let $\mathcal{A}(G)$ denote k-BPDS-TO-MSLR applied with the parameters above to a bipartite graph G with m left vertices and n right vertices. Then $\mathcal{A}$ runs in $\mathrm{poly}(m,n)$ *time and it follows that*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S\times T,p,q)\right),\ \mathrm{MSLR}_D(N,S,d,\gamma,\epsilon)\right) = O\left(w^{-1}+k_n^{-2}m^{-2}r^{-2t}+m^{3/2}n^{-1/2}\right) + O\left(k_n(4e^{-3})^{n/2k_n}+N^{-1}\right)$$

$$d_{TV}\left(\mathcal{A}\left(\mathrm{Bern}(q)^{\otimes m\times n}\right),\ \left(\mathcal{N}(0,I_d)\otimes\mathcal{N}\left(0,1+\gamma^2\right)\right)^{\otimes N}\right) = O\left(k_n^{-2}m^{-2}r^{-2t}+m^{3/2}n^{-1/2}\right)$$

for all subsets $S\subseteq[m]$ with $|S|=k_m$ and subsets $T\subseteq[n]$ with $|T|=k_n$ and $|T\cap F_i|=1$ for each $1\le i\le k_n$.

The proof of this theorem will be broken into several lemmas for clarity. The following four lemmas analyze the approximate Markov transition properties of Steps 4 and 5 of k-BPDS-TO-MSLR. The first three lemmas establishes a total variation upper bound in the single sample case. The fourth lemma is a simple consequence of the first two and establishes the Markov transition properties for Steps 4 and 5 together.

Lemma 53 (Planted Single Sample Labelling) *Let N be a parameter, $\gamma,\mu'\in(0,1)$, $C>0$ be a constant and $u\in\mathbb{R}^d$ be such that $\|u\|_2=1$ and $4C^2\gamma^2\le(\mu')^2/\log N$. Define the random variables (X,y) and (X',y') where $X,X'\in\mathbb{R}^d$ and $y,y'\in\mathbb{R}$ as follows:*

- *Let $X\sim\mathcal{N}(0,I_d)$ and $\eta\sim\mathcal{N}(0,1)$ be independent, and define*

$$y=\gamma\cdot\langle u,X\rangle+\eta$$

- *Let y' be a sample from $\mathcal{N}(0,1+\gamma^2)$ truncated to satisfy $|y'|\le C\sqrt{(1+\gamma^2)\log N}$, and let $Z\sim\mathcal{N}(\mu'\cdot u,I_d)$, $G\sim\mathcal{N}(0,I_d)$ and $W\sim\mathcal{N}\left(0,I_d-\frac{2\gamma^2}{1+\gamma^2}\cdot uu^\top\right)$ be independent. Now let X' be*

$$X'=\frac{1}{\sqrt{2}}\left(\frac{\gamma\cdot y'\sqrt{2}}{\mu'(1+\gamma^2)}\cdot Z+\sqrt{1-2\left(\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\right)^2}\cdot G+W\right) \tag{8}$$

Then it follows that, as $N\to\infty$,

$$d_{TV}\left(\mathcal{L}(X,y),\mathcal{L}(X',y')\right)=O\left(N^{-C^2/2}\right)$$

Proof First observe that $4C^2\gamma^2\le(\mu')^2/\log N$ implies that since $|y'|\le C\sqrt{(1+\gamma^2)\log N}$ holds almost surely and $\gamma\in(0,1)$, it follows that

$$2\left(\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\right)^2\le 2(1+\gamma^2)C^2\gamma^2(\mu')^{-2}\log N\le 1$$

and hence X' is well-defined almost surely.

Now note that since y is a linear function of X and η, which are independent Gaussians, it follows that the $d+1$ entries of (X,y) are jointly Gaussian. Since $\|u\|_2=1$, it follows that

$\mathrm{Var}(y) = 1 + \gamma^2$ and furthermore $\mathrm{Cov}(y, X) = \mathbb{E}[Xy] = \gamma \cdot u$. This implies that the covariance matrix of (X, y) is given by

$$\begin{bmatrix} I_d & \gamma \cdot u \\ \gamma \cdot u^\top & 1 + \gamma^2 \end{bmatrix}$$

It is well known that $X|y$ is a Gaussian vector with mean and covariance matrix given by

$$\mathcal{L}(X|y) = \mathcal{N}\left(\frac{\gamma \cdot y}{1+\gamma^2} \cdot u,\ I_d - \frac{\gamma^2}{1+\gamma^2} \cdot uu^\top\right)$$

Now consider $\mathcal{L}(X'|y')$. Let $Z = \mu' \cdot u + G'$ where $G' \sim \mathcal{N}(0, I_d)$ and note that

$$X' = \frac{\gamma \cdot y'}{1+\gamma^2} \cdot u + \frac{\gamma \cdot y'}{\mu'(1+\gamma^2)} \cdot G' + \frac{1}{\sqrt{2}} \cdot \sqrt{1 - 2\left(\frac{\gamma \cdot y'}{\mu'(1+\gamma^2)}\right)^2} \cdot G + \frac{1}{\sqrt{2}} \cdot W$$

Note that since y', G', G and W are independent, it follows that all of the entries of the second, third and fourth terms in the expression above are jointly Gaussian conditioned on y'. Therefore the entries of $X'|y'$ are also jointly Gaussian. Furthermore the second, third and fourth terms in the expression above for X' have covariance matrices given by

$$\left(\frac{\gamma \cdot y'}{\mu'(1+\gamma^2)}\right)^2 \cdot I_d, \quad \left(\frac{1}{2} - \left(\frac{\gamma \cdot y'}{\mu'(1+\gamma^2)}\right)^2\right) \cdot I_d \quad \text{and} \quad \frac{1}{2} \cdot I_d - \frac{\gamma^2}{1+\gamma^2} \cdot uu^\top$$

respectively, conditioned on y'. Since these three terms are independent conditioned on y', it follows that $X'|y'$ has covariance matrix $I_d - \frac{\gamma^2}{1+\gamma^2} \cdot uu^\top$ and therefore that

$$\mathcal{L}(X'|y') = \mathcal{N}\left(\frac{\gamma \cdot y}{1+\gamma^2} \cdot u,\ I_d - \frac{\gamma^2}{1+\gamma^2} \cdot uu^\top\right)$$

and is hence identically distributed to $\mathcal{L}(X|y)$. Let $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$ be the CDF of $\mathcal{N}(0,1)$. The conditioning property of total variation in Fact 15 therefore implies that

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathcal{L}(X,y), \mathcal{L}(X',y')\right) &\le d_{\mathrm{TV}}\left(\mathcal{L}(y), \mathcal{L}(y')\right) \\ &= \mathbb{P}\left[|y| > c\sqrt{(1+\gamma^2)\log N}\right] \\ &= 2 \cdot \left(1 - \Phi\left(C\sqrt{\log N}\right)\right) \\ &= O\left(N^{-C^2/2}\right) \end{aligned}$$

where the first equality holds due to the conditioning on an event property of total variation in Fact 15 and the last upper bound follows from the standard estimate $1 - \Phi(x) \le \frac{1}{\sqrt{2\pi}} \cdot x^{-1} \cdot e^{-x^2/2}$ for $x \ge 1$. This completes the proof of the lemma. ■

The next lemma establishes single sample guarantees that will be needed to analyze the case in which $\epsilon < 1/2$. The proof of this lemma is very similar to that of Lemma 53 and is deferred to Appendix Q.4.

Lemma 54 (Imbalanced Planted Single Sample Labelling) *Let N, γ, μ', C and u be as in Lemma 53 and let $\mu'' \in (0,1)$. Define the random variables (X, y) and (X', y') as follows:*

- *Let (X, y) where $X \in \mathbb{R}^d$ and $y \in \mathbb{R}$ be jointly Gaussian with mean zero and $(d+1)\times(d+1)$ covariance matrix given by*

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{Xy} \\ \Sigma_{yX} & \Sigma_{yy} \end{bmatrix} = \begin{bmatrix} I_d + \frac{(a^2-1)\gamma^2}{1+\gamma^2} \cdot uu^\top & a\gamma \cdot u \\ a\gamma \cdot u^\top & 1+\gamma^2 \end{bmatrix}$$

- *Let y', Z, G and W be independent where y', G and W are distributed as in Lemma 53 and $Z \sim \mathcal{N}(\mu'' \cdot u, I_d)$. Let X' be defined by Equation (8) as in Lemma 53.*

Then it follows that, as $N \to \infty$,

$$d_{TV}\left(\mathcal{L}(X, y), \mathcal{L}(X', y')\right) = O\left(N^{-C^2/2}\right)$$

We now state a similar lemma analyzing a single sample in Step 4 of k-BPDS-TO-MSLR in the case where X and W are not planted. Its proof is also deferred to Appendix Q.4.

Lemma 55 (Unplanted Single Sample Labelling) *Let N, γ, μ', C and u be as in Lemma 53. Suppose that y' is a sample from $\mathcal{N}(0, 1+\gamma^2)$ truncated to satisfy $|y'| \le C\sqrt{(1+\gamma^2)\log N}$ and $Z, G, W \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$ are independent. Let X' be defined by Equation (8) as in Lemma 53. Then, as $N \to \infty$,*

$$d_{TV}\left(\mathcal{L}(X', y'), \mathcal{N}(0, I_d) \otimes \mathcal{N}(0, 1+\gamma^2)\right) = O\left(N^{-C^2/2}\right)$$

Combining these three lemmas, we now can analyze Step 4 and Step 5 of $\mathcal{A}$. Let $\mathcal{A}_{\text{4-5}}(Z, W)$ denote Steps 4 and 5 of $\mathcal{A}$ with inputs $Z = (Z_1, Z_2, \ldots, Z_N)$ and $W = (W_1, W_2, \ldots, W_n)$ and output $((X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N))$. The next lemma applies the previous two lemmas to establish the Markov transition properties of $\mathcal{A}_{\text{4-5}}$.

Lemma 56 (Scaling and Labelling ISGM Instances) *Let $r, N, d, \gamma, \epsilon, m, n, k_n, k_m$ and $S \subseteq [m]$ where $|S| = k_m$ be as in Theorem 52 and let $\mu, \gamma, \theta > 0$ be such that*

$$\mu = 4\gamma \cdot \sqrt{\frac{\log N}{k_m}}, \quad \tau^2 = \frac{8n\gamma^2}{k_n k_m (1-\gamma^2)} \quad \text{and} \quad \theta = \frac{\tau^2 k_n k_m}{4n + \tau^2 k_n k_m}$$

If $Z \sim \text{ISGM}(N, S, d, \mu, \epsilon)$ and $W \sim \mathcal{N}\left(0,\ I_d - \theta v_S v_S^\top\right)^{\otimes n}$, then

$$d_{TV}\left(\mathcal{A}_{\text{4-5}}(Z, W),\ \text{MSLR}_D(N, S, d, \gamma, \epsilon)\right) = O\left(N^{-1}\right)$$

If $Z \sim \mathcal{N}(0, I_d)^{\otimes N}$ and $W \sim \mathcal{N}(0,1)^{\otimes d \times n}$, then

$$d_{TV}\left(\mathcal{A}_{\text{4-5}}(Z, W),\ \left(\mathcal{N}(0, I_d) \otimes \mathcal{N}(0,1)\right)^{\otimes N}\right) = O\left(N^{-1}\right)$$

Proof We treat the cases in which $\epsilon = 1/2$ and $\epsilon < 1/2$ as well as the two possible distributions of (Z, W) in the lemma statement separately. We first consider the case where $\epsilon = 1/2$ and $r = 2$ and $Z \sim \text{ISGM}_D(N, S, d, \mu, \epsilon)$ and $W \sim \mathcal{N}\left(0,\, I_d - \theta v_S v_S^\top\right)^{\otimes n}$. The Z_i are independent and can be generated by first sampling $s_1, s_2, \ldots, s_N \sim_{\text{i.i.d.}} \text{Bern}(1/2)$ and then setting

$$Z_i \sim \begin{cases} \mathcal{N}(\mu\sqrt{k_m} \cdot v_S, I_d) & \text{if } s_i = 1 \\ \mathcal{N}(-\mu\sqrt{k_m} \cdot v_S, I_d) & \text{if } s_i = 0 \end{cases}$$

where $v_S = k_m^{-1/2} \cdot \mathbf{1}_S$. Let $\mu' = \mu\sqrt{k_m}$. It can be verified that the settings of μ, γ and θ above ensure that

$$\frac{\gamma\sqrt{2}}{\mu'(1+\gamma^2)} = \frac{1}{4(1+\gamma^2)} \cdot \sqrt{\frac{2}{\log N}} \quad \text{and} \quad \theta = \frac{2\gamma^2}{1+\gamma^2}$$

Let $X \sim \mathcal{N}(0, I_d)$ and $\eta \sim \mathcal{N}(0, 1)$ be independent. Applying Lemma 53 with $\mu' = \mu\sqrt{k_m}$, $C = 2$, $u = v_S$ and $u = -v_S$, the equalities above and the definition of X_i in Figure 10 now imply that

$$d_{\text{TV}}\left(\mathcal{L}(X_i, y_i | s_i = 1), \mathcal{L}\left(X, \gamma \cdot \langle v_S, X\rangle + \eta\right)\right) = O(N^{-2})$$
$$d_{\text{TV}}\left(\mathcal{L}(X_i, y_i | s_i = 0), \mathcal{L}\left(X, -\gamma \cdot \langle v_S, X\rangle + \eta\right)\right) = O(N^{-2})$$

for each $1 \le i \le N$. The conditioning property of total variation from Fact 15 now implies that if $\mathcal{L}_1 = \mathcal{L}\left(X, \gamma \cdot \langle v_S, X\rangle + \eta\right)$ and $\mathcal{L}_2 = \mathcal{L}\left(X, -\gamma \cdot \langle v_S, X\rangle + \eta\right)$, then we have that

$$d_{\text{TV}}\left(\mathcal{L}(X_i, y_i), \text{MIX}_{1/2}(\mathcal{L}_1, \mathcal{L}_2)\right) = O(N^{-2})$$

For the given distribution on (Z, W), observe that the pairs (X_i, y_i) for $1 \le i \le N$ are independent by construction in $\mathcal{A}$. Thus the tensorization property of total variation from Fact 15 implies that

$$d_{\text{TV}}\left(\mathcal{L}\left((X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\right),\, \text{MSLR}(N, S, d, \gamma, 1/2)\right) = O(N^{-1})$$

where $\text{MSLR}_D(N, S, d, \gamma, 1/2) = \text{MIX}_{1/2}(\mathcal{L}_1, \mathcal{L}_2)^{\otimes N}$, which establishes the desired bound when $\epsilon = 1/2$ and for the first distribution of (Z, W).

The other two cases will follow by nearly identical arguments. Consider the case where ϵ is arbitrary and if $Z \sim \mathcal{N}(0, I_d)^{\otimes N}$ and $W \sim \mathcal{N}(0, 1)^{\otimes d \times n}$, applying Lemma 55 with $C = 2$ and $\mu' = \mu\sqrt{k_m}$ yields that

$$d_{\text{TV}}\left(\mathcal{L}(X_i, y_i), \mathcal{N}(0, I_d) \otimes \mathcal{N}(0, 1)\right) = O(N^{-2})$$

Applying the tensorization property of total variation from Fact 15 as above then implies the second bound in the lemma statement. Finally, consider the case in which $\epsilon < 1/2$, $r > 2$ and (Z, W) is still distributed as $Z \sim \text{ISGM}_D(N, S, d, \mu, \epsilon)$ and $W \sim \mathcal{N}\left(0,\, I_d - \theta v_S v_S^\top\right)^{\otimes n}$. If the s_i are defined as above, then the Z_i are distributed as

$$Z_i \sim \begin{cases} \mathcal{N}\left(\mu\sqrt{k_m} \cdot v_S, I_d\right) & \text{if } s_i = 1 \\ \mathcal{N}\left(-a\mu\sqrt{k_m} \cdot v_S, I_d\right) & \text{if } s_i = 0 \end{cases}$$

where $a = \epsilon^{-1}(1-\epsilon)$. Now consider applying Lemma 54 with $\mu' = \mu\sqrt{k_m}$, $\mu'' = a\mu' = \mu\epsilon^{-1}(1-\epsilon)$, $C = 2$ and $u = -v_S$. This yields that

$$d_{\text{TV}}\left(\mathcal{L}(X_i, y_i | s_i = 0), \mathcal{L}(X, y)\right) = O(N^{-2})$$

where X and y are as in the statement of Lemma 54. Combining this with the conditioning property of total variation from Fact 15, the application of Lemma 53 in the first case above, the tensorization property of total variation from Fact 15 as in the previous argument and Definition 51 yields that

$$d_{\text{TV}}\left(\mathcal{L}\left((X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\right), \text{MSLR}(N, S, d, \gamma, \epsilon)\right) = O\left(N^{-1}\right)$$

which completes the proof of the lemma. ■

With this lemma, the proof of Theorem 52 reduces to an application of Lemma 16 through a similar argument to the proof of Theorem 46.

Proof [Proof of Theorem 52] Define the steps of $\mathcal{A}$ to map inputs to outputs as follows

$$M \xrightarrow{\mathcal{A}_1} (M_{\text{ISGM}}, M_{\text{NEG-SPCA}}) \xrightarrow{\mathcal{A}_2} (Z, M_{\text{NEG-SPCA}}) \xrightarrow{\mathcal{A}_3} (Z, W) \xrightarrow{\mathcal{A}_{4\text{-}5}} ((X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N))$$

where $Z = (Z_1, Z_2, \ldots, Z_N)$ and $W = (W_1, W_2, \ldots, W_n)$ in Figure 10. First note that the condition on γ in the theorem statement along with the settings of μ and τ in Figure 10 imply that

$$\tau \leq \frac{\delta}{2\sqrt{6\log(mn) + 2\log(p-Q)^{-1}}} \quad \text{where} \quad \delta = \min\left\{\log\left(\frac{p}{Q}\right), \log\left(\frac{1-Q}{1-p}\right)\right\}$$
$$\mu \leq \frac{\delta}{2\sqrt{6\log(k_n m r^t) + 2\log(p-Q)^{-1}}} \cdot \frac{1}{\sqrt{r^t(r-1)(1+(r-1)^{-1})}}$$

for a sufficiently small constant $c > 0$ since $0 < q < p \leq 1$ are constants. Let θ and v_S be as in Lemma 56. Consider Lemma 16 applied to the steps $\mathcal{A}_i$ above and the following sequence of distributions

$$\begin{aligned}
\mathcal{P}_0 &= \mathcal{M}_{[m]\times[n]}(S \times T, \text{Bern}(p), \text{Bern}(q)) \\
\mathcal{P}_1 &= \mathcal{M}_{[m]\times[n]}(S \times T, \text{Bern}(p), \text{Bern}(Q)) \otimes \mathcal{M}_{[m]\times[n]}(S \times T, \text{Bern}(p), \text{Bern}(Q)) \\
\mathcal{P}_2 &= \text{ISGM}_D(N, S, d, \mu, \epsilon) \otimes \mathcal{M}_{[m]\times[n]}(S \times T, \text{Bern}(p), \text{Bern}(Q)) \\
\mathcal{P}_3 &= \text{ISGM}_D(N, S, d, \mu, \epsilon) \otimes \mathcal{N}\left(0, I_d - \theta v_S v_S^\top\right)^{\otimes n} \\
\mathcal{P}_{4\text{-}5} &= \text{MSLR}_D(N, S, d, \gamma, \epsilon)
\end{aligned}$$

Combining the inequalities above for μ and τ with Lemmas 22 and 56 and Theorems 46 and 43 implies that we can take

$$\epsilon_1 = 0, \quad \epsilon_2 = O\left(w^{-1} + k_n^{-2} m^{-2} r^{-2t}\right), \quad \epsilon_3 = O\left(m^{3/2} n^{-1/2} + k_n (4e^{-3})^{n/2k_n}\right) \quad \text{and} \quad \epsilon_{4\text{-}5} = O(N^{-1})$$

Applying Lemma 16 now yields the first total variation upper bound in the theorem. Now consider Lemma 16 applied to

$$\begin{aligned}
\mathcal{P}_0 &= \text{Bern}(q)^{\otimes m\times n} \\
\mathcal{P}_1 &= \text{Bern}(Q)^{\otimes m\times n} \otimes \text{Bern}(Q)^{\otimes m\times n} \\
\mathcal{P}_2 &= \mathcal{N}(0, I_d)^{\otimes N} \otimes \text{Bern}(Q)^{\otimes m\times n} \\
\mathcal{P}_3 &= \mathcal{N}(0, I_d)^{\otimes N} \otimes \mathcal{N}(0, I_d)^{\otimes n}
\end{aligned}$$

$$\mathcal{P}_{\text{4-5}} = \left(\mathcal{N}(0, I_d) \otimes \mathcal{N}(0, 1+\gamma^2)\right)^{\otimes N}$$

By Lemmas 22 and 56 and Theorems 46 and 43, we can take

$$\epsilon_1 = 0, \quad \epsilon_2 = O\left(k_n^{-2} m^{-2} r^{-2t}\right), \quad \epsilon_3 = O\left(m^{3/2} n^{-1/2}\right) \quad \text{and} \quad \epsilon_{\text{4-5}} = O(N^{-1})$$

Applying Lemma 16 now yields the second total variation upper bound in the theorem and completes the proof of the theorem. ■

As in the previous section, the random matrix $R_{L,\epsilon}$ can be used in place of $K_{r,t}$ in our reduction k-BPDS-TO-MSLR. Specifically, replacing k-BPDS-TO-ISGM in Step 2 with k-BPDS-TO-ISGM$_R$ and again replacing r^t with the more flexible parameter L yields an alternative reduction k-BPDS-TO-MSLR$_R$. The guarantees below for this modified reduction follow from the same argument as in the proof of Theorem 52, using Corollary 49 in place of Theorem 46.

Corollary 57 (Reduction from k-BPDS to MSLR with $R_{L,\epsilon}$) *Let n be a parameter and let $w(n) = \omega(1)$ be a slow-growing function. Fix initial and target parameters as follows:*

- Initial k-BPDS Parameters: *m, n, k_m, k_n, p, q and F as in Theorem 52.*
- Target MSLR Parameters: *(N, d, γ, ϵ) and a parameter $L = L(N) \in \mathbb{N}$ such that $N \le n$ and (N, d, ϵ, L) satisfy the conditions in Corollary 49. Suppose that $\gamma \in (0, 1/2)$ satisfies that*

$$\gamma^2 \le c \cdot \min\left\{\frac{\epsilon k_m}{L \log(k_n m L) \log N}, \frac{k_n k_m}{n \log(mn)}\right\}$$

  for a sufficiently small constant $c > 0$.

If $\mathcal{A}$ denotes k-BPDS-TO-MSLR$_R$ applied with the parameters above, then $\mathcal{A}$ runs in $\text{poly}(m, n)$ *time and*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S \times T, p, q)\right), \text{MSLR}_D(N, S, d, \gamma, \epsilon)\right) = o(1)$$
$$d_{TV}\left(\mathcal{A}\left(\text{Bern}(q)^{\otimes m \times n}\right), \left(\mathcal{N}(0, I_d) \otimes \mathcal{N}\left(0, 1+\gamma^2\right)\right)^{\otimes N}\right) = o(1)$$

for all k_m-subsets $S \subseteq [m]$ and k_n-subsets $T \subseteq [n]$ with $|T \cap F_i| = 1$ for each $1 \le i \le k_n$.

Appendix J. Completing Tensors from Hypergraphs

In this section we introduce a key subroutine that will be used in our reduction to tensor PCA in Section N. The starting point for our reduction k-HPDS-TO-TPCA is the hypergraph problem k-HPDS. The adjacency tensor of this instance is missing all entries with at least one pair of equal indices. The first procedure ADVICE-COMPLETE-TENSOR in this section gives a method of completing these missing entries and producing an instance of the planted sub-tensor problem, given access to a set of $s-1$ vertices in the clique, where s is the order of the target tensor. In order to translate this into a reduction, we iterate over all $(s-1)$-sets of vertices and carry out this reduction for each one, as will be described in more detail later in this section. For the motivation and high-level ideas behind the reductions in this section, we refer to the discussion in Section C.6.

In order to describe our reduction ADVICE-COMPLETE-TENSOR, we will need the following definition which will be crucial in indexing the missing entries of the tensor.

Algorithm ADVICE-COMPLETE-TENSOR

Inputs: HPDS instance $H \in \mathcal{G}_n^s$ with edge probabilities $0 < q < p \leq 1$, an $(s-1)$-set of advice vertices $V = \{v_1, v_2, \ldots, v_{s-1}\}$ of H

1. *Clone Hyperedges*: Compute the $(s!)^2$ hypergraphs $H^{\sigma_1,\sigma_2} \in \mathcal{G}_n^s$ for each pair $\sigma_1, \sigma_2 \in S_s$ by applying BERNOULLI-CLONE with $t = (s!)^2$ to the $\binom{N}{s}$ hyperedge indicators of H with input Bernoulli probabilities p and q and output probabilities p and

$$Q = 1 - (1-p)^{1-1/t}(1-q)^{1/t} + \mathbf{1}_{\{p=1\}}\left(q^{1/t} - 1\right)$$

2. *Form Tensor Entries*: For each $I = (i_1, i_2, \ldots, i_s) \in ([N]\backslash V)^s$, set the $(i_1, i_2, \ldots, i_s)$th entry of the tensor T with dimensions $(N-s+1)^{\otimes s}$ to be the following hyperedge indicator

$$T_{i_1,i_2,\ldots,i_s} = \mathbf{1}\left\{\{v_1, v_2, \ldots, v_{s-|P(I)|}\} \cup \{i_1, i_2, \ldots, i_s\} \in E\left(H^{\tau_{\mathrm{P}}(I),\tau_{\mathrm{V}}(I)}\right)\right\}$$

   where $P(I)$, $\tau_{\mathrm{P}}(I)$ and $\tau_{\mathrm{V}}(I)$ are as in Definition 58.

3. *Output*: Output the order s tensor T with axes indexed by the set $[N]\backslash V$.

Algorithm ITERATE-AND-REDUCE

Inputs: k-HPDS instance $H \in \mathcal{G}_n^s$ with edge probabilities $0 < q < p \leq 1$, partition E of $[n]$ into k equally-sized parts, a one-sided blackbox $\mathcal{B}$ for the corresponding planted tensor problem

1. For every $(s-1)$-set of vertices $\{v_1, v_2, \ldots, v_{s-1}\}$ all from different parts of E, form the tensor T by applying ADVICE-COMPLETE-TENSOR to H and $\{v_1, v_2, \ldots, v_{s-1}\}$, remove the indices of T that are in the same part of E as at least one of $\{v_1, v_2, \ldots, v_{s-1}\}$ and run the blackbox $\mathcal{B}$ on the resulting tensor T.

2. Output H_1 if any application if $\mathcal{B}$ in Step 1 output H_1.

Figure 11: The first reduction is a subroutine to complete the entries of a planted dense sub-hypergraph problem into a planted tensor problem given an advice set of vertices. The second reduction uses this subroutine to reduce solving a planted dense sub-hypergraph problem to producing a one-sided blackbox solving the planted tensor problem.

Definition 58 (Tuple Statistics) *Given a tuple $I = (i_1, i_2, \ldots, i_s)$ where each $i_j \in U$ for some set U, we define the partition $P(I)$ and permutations $\tau_{\mathrm{P}}(I)$ and $\tau_{\mathrm{V}}(I)$ of $[s]$ as follows:*

1. *Let $P(I)$ be the unique partition of $[s]$ into nonempty parts $P_1, P_2, \ldots, P_t$ where $i_k = i_l$ if and only if $k, l \in P_j$ for some $1 \leq j \leq t$, and let $|P(I)| = t$.*

2. *Given the partition $P(I)$, let $\tau_{\mathrm{P}}(I)$ be the permutation of $[s]$ formed by ordering the parts P_j in increasing order of their largest element, and then listing the elements of the parts P_j*

according to this order, where the elements of each individual part are written in decreasing order.

3. *Let $P'_1, P'_2, \ldots, P'_t$ be the ordering of the parts of $P(I)$ as defined above and let $v_1, v_2, \ldots, v_t$ be such that $v_j = i_k$ for all $k \in P'_j$ or in other words v_j is the common value of i_k of all indices k in the part P'_j. The values $v_1, v_2, \ldots, v_t$ are by definition distinct and their ordering induces a permutation σ on $[t]$. Let $\tau_{\mathrm{V}}(I)$ be the permutation on $[s]$ formed by setting $(\tau_{\mathrm{V}}(I))_{[t]} = \sigma$ and extending σ to $[s]$ by taking $(\tau_{\mathrm{V}}(I))(j) = j$ for all $t < j \leq s$.*

Note that $|P(I)|$ is the number of distinct values in I and thus $|P(I)| = |\{i_1, i_2, \ldots, i_s\}|$ for each I. For example, if $I = (4, 4, 1, 2, 2, 5, 3, 5, 2)$ and $s = 9$, then $P(I)$, $\tau_{\mathrm{P}}(I)$ and $\tau_{\mathrm{V}}(I)$ are

$$P(I) = \{\{1, 2\}, \{3\}, \{4, 5, 9\}, \{6, 8\}, \{7\}\}, \quad \tau_{\mathrm{P}}(I) = (2, 1, 3, 7, 8, 6, 9, 5, 4) \quad \text{and}$$
$$\tau_{\mathrm{V}}(I) = (4, 1, 3, 5, 2, 6, 7, 8, 9)$$

We now establish the main Markov transition properties of ADVICE-COMPLETE-TENSOR. Given a set X, let $\mathcal{E}_{X,s}$ be the set $\binom{X}{s}$ of all subsets of X of size s.

Lemma 59 (Completing Tensors with Advice Vertices) *Let $0 < q < p \leq 1$ be such that $\min\{q, 1-q\} = \Omega_N(1)$ and let s be a constant. Let $0 < Q < p$ be given by*

$$Q = 1 - (1-p)^{1-1/t}(1-q)^{1/t} + \mathbf{1}_{\{p=1\}}\left(q^{1/t} - 1\right)$$

where $t = (s!)^2$. Let V be an arbitrary $(s-1)$-subset of $[N]$ and let $\mathcal{A}$ denote ADVICE-COMPLETE-TENSOR *with input H, output T, advice vertices V and parameters p and q. Then $\mathcal{A}$ runs in* $\mathrm{poly}(N)$ *time and satisfies*

$$\mathcal{A}\left(\mathcal{M}_{\mathcal{E}_{[N],s}}\left(\mathcal{E}_{S\cup V,s}, \mathrm{Bern}(p), \mathrm{Bern}(q)\right)\right) \sim \mathcal{M}_{([N]\setminus V)^s}\left(S^s, \mathrm{Bern}(p), \mathrm{Bern}(Q)\right)$$
$$\mathcal{A}\left(\mathcal{M}_{\mathcal{E}_{[N],s}}\left(\mathrm{Bern}(q)\right)\right) \sim \mathcal{M}_{([N]\setminus V)^s}\left(\mathrm{Bern}(Q)\right)$$

for all subsets $S \subseteq [N]$ disjoint from V.

Proof First note that Step 2 of $\mathcal{A}$ is well defined since the fact that $|P(I)| = |\{i_1, i_2, \ldots, i_s\}|$ implies that $\{v_1, v_2, \ldots, v_{s-|P(I)|}\} \cup \{i_1, i_2, \ldots, i_s\}$ is always a set of size s. We first consider the case in which $H \sim \mathcal{M}_{\mathcal{E}_{[N],s}}\left(\mathcal{E}_{S\cup V,s}, \mathrm{Bern}(p), \mathrm{Bern}(q)\right)$. By Lemma 22, it follows that the hyperedge indicators of H^{σ_1,σ_2} are all independent and distributed as

$$\mathbf{1}\left\{e \in E\left(H^{\sigma_1,\sigma_2}\right)\right\} \sim \begin{cases} \mathrm{Bern}(p) & \text{if } e \subseteq S \cup V \\ \mathrm{Bern}(Q) & \text{otherwise} \end{cases}$$

for each $\sigma_1, \sigma_2 \in S_s$ and subset $e \subseteq [N]$ with $|e| = s$. We now observe that T agrees in its entrywise marginal distributions with $\mathcal{M}_{([N]\setminus V)^s}\left(S^s, \mathrm{Bern}(p), \mathrm{Bern}(Q)\right)$. In particular, we have that:

- if $(i_1, i_2, \ldots, i_s)$ is such that $i_j \in S$ for all $1 \leq j \leq s$ then we have that $\{v_1, v_2, \ldots, v_{s-|P(I)|}\} \cup \{i_1, i_2, \ldots, i_s\} \subseteq S \cup V$ and hence

$$T_{i_1,i_2,\ldots,i_s} = \mathbf{1}\left\{\{v_1, v_2, \ldots, v_{s-|P(I)|}\} \cup \{i_1, i_2, \ldots, i_s\} \in E\left(H^{\tau_{\mathrm{P}}(I),\tau_{\mathrm{V}}(I)}\right)\right\} \sim \mathrm{Bern}(p)$$

- if $(i_1, i_2, \ldots, i_s)$ is such that there is some j such that $i_j j \notin S$, then $\{v_1, v_2, \ldots, v_{s-|P(I)|}\} \cup \{i_1, i_2, \ldots, i_s\} \not\subseteq S \cup V$ and $T_{i_1,i_2,\ldots,i_s} \sim \text{Bern}(Q)$.

It suffices to verify that the entries of T are independent. Since all of the hyperedge indicators of the H^{σ_1,σ_2} are independent, it suffices to verify that the entries of T are equal to distinct hyperedge indicators.

To show this, we will show that $\{i_1, i_2, \ldots, i_s\}$, $\tau_{\text{P}}(I)$ and $\tau_{\text{V}}(I)$ determine the tuple $I = (i_1, i_2, \ldots, i_s)$, from which the desired result follows. Consider the longest increasing subsequence of $\tau_{\text{P}}(I)$ starting with $(\tau_{\text{P}}(I))\,(1)$. The elements of this subsequence partition $\tau_{\text{P}}(I)$ into contiguous subsequences corresponding to the parts of $P(I)$. Thus $\tau_{\text{P}}(I)$ determines $P(I)$. Now the first $|P(I)|$ elements of $\tau_{\text{V}}(I)$ along with $\{i_1, i_2, \ldots, i_s\}$ determine the values v_j in Definition 58 corresponding to I on each part of $P(I)$. This uniquely determines the tuple I. Therefore the entries $T_{i_1,i_2,\ldots,i_s}$ all correspond to distinct hyperedge indicators and are therefore independent. Applying this argument with $S = \emptyset$ yields the second identity in the statement of the lemma. This completes the proof of the lemma. ■

We now analyze the additional subroutine ITERATE-AND-REDUCE. This will show it suffices to design a reduction with low total variation error in order to show computational lower bounds for Tensor PCA. Let $k\text{-PST}_E^s(N, k, p, q)$ denote the following *planted subtensor* hypothesis testing problem with hypotheses

$$H_0 : T \sim \mathcal{M}_{[N]^s}\left(\text{Bern}(q)\right) \quad \text{and} \quad H_1 : T \sim \mathcal{M}_{[N]^s}\left(S^s, \text{Bern}(p), \text{Bern}(q)\right)$$

where S is chosen uniformly at random from all k-subsets of $[N]$ intersecting each part of E in one element. The next lemma captures our key guarantee of ITERATE-AND-REDUCE.

Lemma 60 (Hardness of One-Sided Blackboxes by Reduction) *Fix a pair $0 < q < p \le 1$ with $\min\{q, 1-q\} = \Omega(1)$, a constant s and let Q be as in Figure 11. Suppose that there is a reduction mapping both hypotheses of $k\text{-PST}_E^s(N - (s-1)N/k, k - s + 1, p, Q)$ with $k = o(\sqrt{N})$ to the corresponding hypotheses H_0 and H_1 of a testing problem $\mathcal{P}$ within total variation $O(N^{-s})$. Then the $k\text{-HPC}^s$ or $k\text{-HPDS}^s$ conjecture for constant $0 < q < p \le 1$ implies that there cannot be a $\text{poly}(n)$ time algorithm $\mathcal{A}$ solving $\mathcal{P}$ with a low false positive probability of $\mathbb{P}_{H_0}[\mathcal{A}(X) = H_1] = O(N^{-s})$, where X denotes the observed variable in $\mathcal{P}$.*

Proof Assume for contradiction that there is a such a $\text{poly}(n)$ time algorithm $\mathcal{A}$ for $\mathcal{P}$ with $\mathbb{P}_{H_0}[\mathcal{A}(X) = H_1] = O(N^{-s})$ and Type I+II error

$$\mathbb{P}_{H_0}[\mathcal{A}(X) = H_1] + \mathbb{P}_{H_1}[\mathcal{A}(X) = H_0] \le 1 - \epsilon$$

for some $\epsilon = \Omega(1)$. Furthermore, let $\mathcal{R}$ denote the reduction described in the lemma. If H_0' and H_1' denote the hypotheses of $k\text{-PST}_E^s(N - (s-1)N/k, k - s + 1, p, Q)$ and T denotes an instance of this problem, then $\mathcal{R}$ satisfies that

$$d_{\text{TV}}\left(\mathcal{R}\left(\mathcal{L}_{H_0'}(T)\right), \mathcal{L}_{H_0}(T)\right) + d_{\text{TV}}\left(\mathcal{R}\left(\mathcal{L}_{H_1'}(T)\right), \mathcal{L}_{H_1}(T)\right) = O(N^{-s})$$

Now consider applying ITERATE-AND-REDUCE to: (1) a hard instance H of $k\text{-HPDS}(N, k, p, q)$ with $k = o(\sqrt{N})$; and (2) the blackbox $\mathcal{B} = \mathcal{A} \circ \mathcal{R}$. Let $\text{IR}(H) \in \{H_0'', H_1''\}$ denote the output of

ITERATE-AND-REDUCE on input H, and let H_0'' and H_1'' be the hypotheses of k-HPDS(N, k, p, q). Furthermore, let $T_1, T_2, \ldots, T_K$ denote the tensors formed in the $K = \left(\frac{N}{k}\right)^{s-1}\binom{k}{s-1}$ iterations of Step 1 of ITERATE-AND-REDUCE. Note that each T_i has all of its s dimensions equal to $N - (s-1)N/k$ since exactly $s-1$ parts of E of size N/k are removed from $[N]$ in each iteration of Step 1 of ITERATE-AND-REDUCE. First consider the case in which H_0'' holds. Each tensor in the sequence $T_1, T_2, \ldots, T_K$ is marginally distributed as $\mathcal{M}_{[N-(s-1)N/k]^s}(\text{Bern}(Q))$ by Lemma 59. By definition $\text{IR}(H) = H_1''$ if and only if some application of $\mathcal{B}(T_i)$ outputs H_1. Now note that by a union bound, the definition of d_{TV} and the data-processing inequality, we have that

$$\begin{aligned}
\mathbb{P}_{H_0''}\left[\text{IR}(H) = H_1''\right] &\le \sum_{i=1}^{K} \mathbb{P}_{H_0''}[\mathcal{A} \circ \mathcal{R}(T_i) = H_1] \\
&\le \sum_{i=1}^{K} \left[\mathbb{P}_{H_0}[\mathcal{A}(X) = H_1] + d_{\text{TV}}\left(\mathcal{R}\left(\mathcal{L}_{H_0'}(T)\right), \mathcal{L}_{H_0}(T)\right)\right] \\
&= O\left(K \cdot N^{-s}\right) = O(N^{-1})
\end{aligned}$$

since $K = O(N^{s-1})$. Now suppose that H_1'' holds and let i^* be the first iteration of ITERATE-AND-REDUCE in which each of the vertices $\{v_1, v_2, \ldots, v_{s-1}\}$ are in the planted dense sub-hypergraph of H. Lemma 59 shows that T_{i^*} is distributed as $\mathcal{M}_{[N-(s-1)N/k]^s}(S^s, \text{Bern}(p), \text{Bern}(Q))$ where S is chosen uniformly at random over all $(k-s+1)$-subsets of $[N-(s-1)N/k]$ with one element per part of the input partition E associated with H. We now have that

$$\begin{aligned}
\mathbb{P}_{H_1''}\left[\text{IR}(H) = H_0''\right] &\le 1 - \mathbb{P}_{H_1''}\left[\text{IR}(H) = H_1''\right] \le 1 - \mathbb{P}_{H_1''}[\mathcal{A} \circ \mathcal{R}(T_{i^*}) = H_1] \\
&\le 1 - \mathbb{P}_{H_1}[\mathcal{A}(X) = H_1] + d_{\text{TV}}\left(\mathcal{R}\left(\mathcal{L}_{H_1'}(T)\right), \mathcal{L}_{H_1}(T)\right) \\
&= \mathbb{P}_{H_1}[\mathcal{A}(X) = H_0] + O(N^{-s})
\end{aligned}$$

Therefore the Type I+II error of ITERATE-AND-REDUCE is

$$\mathbb{P}_{H_0''}\left[\text{IR}(H) = H_1''\right] + \mathbb{P}_{H_1''}\left[\text{IR}(H) = H_0''\right] = \mathbb{P}_{H_1}[\mathcal{A}(X) = H_0] + O(N^{-1}) \le 1 - \epsilon + O(N^{-1})$$

and ITERATE-AND-REDUCE solves k-HPDS, contradicting the k-HPDS conjecture. ■

Part III
Computational Lower Bounds from PC_ρ

Appendix K. Secret Leakage and Hardness Assumptions

In this section, we further discuss the conditions in the PC_ρ conjecture and provide evidence for it and for the specific hardness assumptions we use in our reductions. In Section K.1, we show that k-HPC^s is our strongest hardness assumption, explicitly give the ρ corresponding to each of these hardness assumptions and show that the barriers in Conjecture 3 are supported by the PC_ρ conjecture for these ρ. In Section K.2, we give more general evidence for the PC_ρ conjecture through the failure of low-degree polynomial tests. We also discuss technical conditions in variants of the low-degree conjecture and how these relate to the PC_ρ conjecture. Finally, in Section K.3, we give evidence supporting several of the barriers in Conjecture 3 from statistical query lower bounds.

We remark that, as mentioned at the end of Section 2, all of our results and conjectures for PC_ρ appear to also hold for PDS_ρ at constant edge densities $0 < q < p \leq 1$. Evidence for these extensions to PDS_ρ from the failure of low-degree polynomials and SQ algorithms can be obtained through computations analogous to those in Sections K.2 and K.3.

K.1. Hardness Assumptions and the PC_ρ Conjecture

In this section, we continue the discussion of the PC_ρ conjecture from Section 2. We first show that k-HPC^s reduces to the other conjectured barriers in Conjecture 3. We then formalize the discussion in Section 2 and explicitly construct secret leakage distributions ρ such that the graph problems in Conjecture 3 can be obtained from instances of PC_ρ with these ρ. We then verify that the PC_ρ conjecture implies Conjecture 3 up to arbitrarily small polynomial factors. More precisely, we verify that these ρ, when constrained to be in the conjecturally hard parameter regimes in Conjecture 3, satisfy the tail bound conditions on $p_\rho(s)$ in the PC_ρ conjecture.

The k-HPC^s Conjecture is the Strongest Hardness Assumption. First note that when $s = 2$, our conjectured hardness for k-HPC^s is exactly our conjectured hardness for k-PC in Conjecture 3. Thus it suffices to show that Conjecture 3 for k-HPC^s implies the conjecture for k-BPC and BPC. This is the content of the following lemma.

Lemma 61 *Let α be a fixed positive rational number and $w = w(n)$ be an arbitrarily slow-growing function with $w(n) \to \infty$. Then there is a positive integer s and a $\text{poly}(n)$ time reduction from k-$\text{HPC}^s(n, k, 1/2)$ with $k = o(\sqrt{n})$ to either of the problems k-$\text{BPC}(M, N, k_M, k_N, 1/2)$ or $\text{BPC}(M, N, k_M, k_N, 1/2)$ for some parameters satisfying $M = \Theta(N^\alpha)$ and $Cw^{-1}\sqrt{N} \leq k_N = o(\sqrt{N})$ and $Cw^{-1}\sqrt{M} \leq k_M = o(\sqrt{M})$ for some positive constant $C > 0$.*

Proof We first describe the desired reduction to k-BPC. Let $\alpha = a/b$ for two fixed integers a and b, and let H be an input instance of k-$\text{HPC}^{a+b}_E(n, k, 1/2)$ where E is a fixed known partition of $[n]$. Suppose that H is a nearly tight instance with $w^{-1/\max(a,b)}\sqrt{n} \leq k = o(\sqrt{n})$. Now consider the following reduction:

1. Let $R_1, R_2, \ldots, R_{a+b}$ be a partition of $[k]$ into $a + b$ sets of sizes differing by at most 1, and let $E(R_j) = \bigcup_{i \in R_j} E_i$ for each $j \in [a + b]$.

2. Form the bipartite graph G with left vertex set indexed by $V_1 = E(R_1) \times E(R_2) \times \cdots \times E(R_a)$ and right vertex set $V_2 = E(R_{a+1}) \times E(R_{a+2}) \times \cdots \times E(R_{a+b})$ such that $(u_1, u_2, \ldots, u_a) \in V_1$ and $(v_1, v_2, \ldots, v_b) \in V_2$ are adjacent if and only if $\{u_1, \ldots, u_a, v_1, \ldots, v_b\}$ is a hyperedge of H.

3. Output G with left parts $E_{i_1} \times E_{i_2} \times \cdots \times E_{i_a}$ for all $(i_1, i_2, \ldots, i_a) \in R_1 \times R_2 \times \cdots \times R_a$ and right parts $E_{i_1} \times E_{i_2} \times \cdots \times E_{i_b}$ for all $(i_1, i_2, \ldots, i_b) \in R_{a+1} \times R_{a+2} \times \cdots \times R_{a+b}$, after randomly permuting the vertex labels of G within each of these parts.

Note that since $a + b = \Theta(1)$, we have that $|E(R_i)| = \Theta(n)$ for each i and thus $N = |V_2| = \Theta(n^b)$ and $M = |V_1| = \Theta(n^a) = \Theta(N^\alpha)$. Under H_0, each possible hyperedge of H is included independently with probability $1/2$. Since the edge indicators of G corresponds to a distinct hyperedge indicator of H in Step 2 above, it follows that each edge of G is also included with probability $1/2$ and thus $G \sim \mathcal{G}_B(M, N, 1/2)$.

In the case of H_1, suppose that H is distributed according to the hypergraph planted clique distribution with clique vertices $S \subseteq [n]$ where $S \sim \mathcal{U}_n(E)$. Examining the definition of the edge indicators in Step 2 above yields that G is a sample from H_1 of k-BPC$(M, N, k_M, k_N, 1/2)$ conditioned on having left biclique set $\prod_{i=1}^{a}(S \cap E(R_i))$ and right biclique set $\prod_{i=a+1}^{a+b}(S \cap E(R_i))$. Observe that these sets have exactly one vertex in G in common with each of the parts described in Step 3 above. Now note that since S has one vertex per part of E, we have that $|S \cap E(R_i)| = |R_i| = \Theta(k)$ since $a + b = \Theta(1)$. Thus $k_M = |\prod_{i=1}^{a}(S \cap E(R_i))| = \Theta(k^a)$ and $k_N = \Theta(k^b)$. The bound on k now implies that the two desired bounds on k_N and k_M hold for a sufficiently small constant $C > 0$. Thus the permutations in Step 3 produce a sample exactly from k-BPC$(M, N, k_M, k_N, 1/2)$ in the desired parameter regime. If instead of only permuting vertex labels within each part, we randomly permute all left vertex labels and all right vertex labels in Step 3, the resulting reduction produces BPC instead of k-BPC. The correctness of this reduction follows from the same argument as for k-BPC. ■

We remark that since m and n are polynomial in each other in the setup in Conjecture 3 for k-BPC and BPC, the lemma above fills out a dense subset of this entire parameter regime – where $m = \Theta(n^\alpha)$ for some rational α. In the case where α is irrational, the reduction in Lemma 61, when composed with our other reductions beginning with k-BPC and BPC, shows tight computational lower bounds up to arbitrarily small polynomial factors n^ϵ by approximating α arbitrarily closely with a rational number.

Hardness Conjectures as Instances of PC$_\rho$. We now will verify that each of the graph problems in Conjecture 3 can be obtained from PC$_\rho$. To do this, we explicitly construct several ρ and give simple reductions from the corresponding instances of PC$_\rho$ to these graph problems. We begin with k-PC, BPC and k-BPC as their discussion will be brief.

Secrets for k-PC, BPC and k-BPC. Below are the ρ corresponding to these three graph problems. Both BPC and k-BPC can be obtained by restricting to bipartite subgraphs of the PC$_\rho$ instances with these ρ.

- **k-partite PC:** Suppose that k divides n and E is a partition of $[n]$ into k parts of size n/k. By definition, k-PC$_E(n, k, 1/2)$ is PC$_\rho(n, k, 1/2)$ where $\rho = \rho_{k\text{-PC}}(E, n, k)$ is the uniform distribution $\mathcal{U}_n(E)$ over all k-sets of $[n]$ intersecting each part of E in one element.

- **bipartite PC:** Let $\rho_{\text{BPC}}(m, n, k_m, k_n)$ be the uniform distribution over all $(k_n + k_m)$-sets of $[n+m]$ with k_n elements in $\{1, 2, \ldots, n\}$ and k_m elements in $\{n+1, n+2, \ldots, n+m\}$. An instance of $\text{BPC}(m, n, k_m, k_n, 1/2)$ can then be obtained by outputting the bipartite subgraph of $\text{PC}_\rho(m + n, k_m + k_n, 1/2)$ with this ρ, consisting of the edges between left vertex set $\{n + 1, n + 2, \ldots, n + m\}$ and right vertex set $\{1, 2, \ldots, n\}$.

- **k-part bipartite PC:** Suppose that k_n divides n, k_m divides m, and E and F are partitions of $[n]$ and $[m]$ into k_n and k_m parts of equal size, respectively. Let $\rho_{k\text{-BPC}}(E, F, m, n, k_m, k_n)$ be uniform over all $(k_n + k_m)$-subsets of $[n + m]$ with exactly one vertex in each part of both E and $n + F$. Here, $n + F$ denotes the partition of $\{n + 1, n + 2, \ldots, n + m\}$ induced by shifting indices in F by n. As with BPC, $k\text{-BPC}(m, n, k_m, k_n, 1/2)$ can be realized as the bipartite subgraph of $\text{PC}_\rho(m + n, k_m + k_n, 1/2)$, with this ρ, between the vertex sets $\{n + 1, n + 2, \ldots, n + m\}$ and $\{1, 2, \ldots, n\}$.

Secret for k-HPC^s. We first will give the secret ρ corresponding to k-HPC^s for even s, which can be viewed as roughly the pushforward of $\mathcal{U}_n(E)$ after unfolding the adjacency tensor of k-HPC^s. The secret for odd s will then be obtained through a slight modification of the even case.

Suppose that $s = 2t$. Given a set $S \subseteq [n]$, let $P_t^n(S)$ denote the subset of $[n^t]$ given by

$$P_t^n(S) = \left\{ 1 + \sum_{j=0}^{t-1} (a_j - 1)n^j : a_0, a_1, \ldots, a_{t-1} \in S \right\}$$

In other words, $P_t^n(S)$ is the set of all numbers x in $[n^t]$ such that the base-n representation of $x - 1$ only has digits in $S - 1$, where $S - 1$ is the set of all $s - 1$ where $s \in S$. Note that if $|S| = k$ then $|P_t^n(S)| = k^t$. Given a partition E of $[n]$ into k parts of size n/k, let $\rho_{k\text{-HPC}^s}(E, n, k)$ be the distribution over k^t-subsets of $[n^t]$ sampled by choosing S at random from $\mathcal{U}_n(E)$ and outputting $P_t^n(S)$. Throughout the rest of this section, we will let $I(a_0, a_1, \ldots, a_{t-1})$ denote the sum $1 + \sum_{j=0}^{t-1}(a_j - 1)n^j$. We now will show that $k\text{-HPC}_E^s(n, k, 1/2)$ can be obtained from $\text{PC}_\rho(n^t, k^t, 1/2)$ where $\rho = \rho_{k\text{-HPC}^s}(E, n, k)$. Intuitively, this instance of PC_ρ has a subset of edges corresponding to the unfolded adjacency tensor of -HPC$_E^s$. More formally, consider the following steps.

1. Let G be an input instance of $\text{PC}_\rho(n^t, k^t, 1/2)$ and let H be the output hypergraph with vertex set $[n]$.

2. Construct H as follows: for each possible hyperedge $e = \{a_1, a_2, \ldots, a_{2t}\}$, with $1 \le a_1 < a_2 < \cdots < a_{2t} \le n$, include e in H if and only if there is an edge between vertices $I(a_1, a_2, \ldots, a_t)$ and $I(a_{t+1}, a_{t+2}, \ldots, a_{2t})$ in G.

Under H_0, it follows that $G \sim \mathcal{G}(n^t, 1/2)$. Note that each hyperedge e in Step 2 identifies a unique pair of distinct vertices $I(a_1, a_2, \ldots, a_t)$ and $I(a_{t+1}, a_{t+2}, \ldots, a_{2t})$ in G, and thus the hyperedges of H are independently included with probability $1/2$. Under H_1, it follows that the instance of $\text{PC}_\rho(n^t, k^t, 1/2)$ is sampled from the planted clique distribution with clique vertices $P_t^n(S)$ where $S \sim \mathcal{U}_n(E)$. By the definition of $P_t^n(S)$, it follows that $I(a_1, a_2, \ldots, a_t)$ is in this clique if and only if $a_1, a_2, \ldots, a_t \in S$. Examining the edge indicators of H then yields that H is a sample from the hypergraph planted clique distribution with clique vertex set S. Since $S \sim \mathcal{U}_n(E)$, under both H_0 and H_1, it follows that H is a sample from k-HPC^s.

Now suppose that s is odd with $s = 2t + 1$. The idea in this case is to pair up adjacent digits in base-n expansions and use these pairs to label the vertices of k-HPC^s. More precisely suppose that $n = N^2$ and $k = K^2$ for some positive integers K and N. Let E be a fixed partition of $[n]$ into $k = K^2$ equally sized parts and let $\rho_{k\text{-HPC}^s}(E, n, k)$ be $\rho_{k\text{-HPC}^{2s}}(F, N, K)$ as defined above for the even number $2s$, where F is a fixed partition of $[N]$ into K equally sized parts. We now will show that k-HPC$^s_E(n, k, 1/2)$ can be obtained from PC$_\rho(N^s, K^s, 1/2)$ where $\rho = \rho_{k\text{-HPC}^{2s}}(F, N, K)$. Let I' be the analogue of I for base-N expansions i.e. let $I(b_0, b_1, \ldots, b_{t-1})$ denote the sum $1 + \sum_{j=0}^{t-1}(b_j - 1)N^j$. Consider the following steps.

1. Let G be an instance of PC$_\rho(N^s, K^s, 1/2)$ and let H be the output hypergraph with vertex set $[n]$.

2. Let $\sigma : [n] \to [n]$ be a bijection such that, for each $i \in [k]$, we have that

$$\sigma(E_i) = \left\{I'(b_0, b_1) : b_0 \in F_{c_0} \text{ and } b_1 \in F_{c_1}\right\}$$

   where c_0, c_1 are the unique elements of $[K]$ with $i - 1 = (c_0 - 1) + (c_1 - 1)K$.

3. Construct H as follows. For each possible hyperedge $e = \{a_1, a_2, \ldots, a_s\}$, with $1 \le a_1 < a_2 < \cdots < a_s \le n$, let b_{2i-1}, b_{2i} be the unique elements of $[N]$ with $I'(b_{2i-1}, b_{2i}) = \sigma(a_i)$ for each i. Now include e in H if and only if there is an edge between the two vertices $I(b_1, b_2, \ldots, b_s)$ and $I(b_{s+1}, b_{s+2}, \ldots, b_{2s})$ in G.

4. Permute the vertex labels of H within each part F_i uniformly at random.

Note that σ always trivially exists because the K^2 sets $E_1, E_2, \ldots, E_{K^2}$ and the K^2 sets $F'_{i,j} = \{I'(b_0, b_1) : b_0 \in F_i \text{ and } b_1 \in F_j\}$ for $1 \le i, j \le K$ are both partitions of $[n]$ into parts of size N^2/K^2. As in the case where s is even, under H_1 we have that $G \sim \mathcal{G}(N^{2s}, 1/2)$ and the hyperedges of H are independently included with probability $1/2$, since Step 3 identifies distinct pairs of vertices for each hyperedge e. Under H_1, let $S \sim \mathcal{U}_N(F)$ be such that the clique vertices in G are $P^N_s(S)$. By the same reasoning as in the even case, after Step 3, the hypergraph H is distributed as a sample from the hypergraph planted clique distribution with clique vertex set $\sigma^{-1}(I'(S, S))$ where $I'(S, S) = \{I'(s_0, s_1) : s_0, s_1 \in S\}$. The definition of σ now ensures that this clique has one vertex per part of E. Step 4 ensures that the resulting hypergraph is exactly a sample from H_1 of k-HPC^s. We remark that the conditions $n = N^2$ and $k = K^2$ do not affect our lower bounds when composing the reduction above with our other reductions. This is due to the subsequence criterion for computational lower bounds in Condition E.1.

Verifying the Conditions of the PC$_\rho$ Conjecture. We now verify that the PC$_\rho$ conjecture corresponds to the hard regimes in Conjecture 3 up to arbitrarily small polynomial factors. To do this, it suffices to verify the tail bound on $p_\rho(s)$ in the PC$_\rho$ conjecture for each ρ described above, which is done in the theorem below. In the next section, we will show that a slightly stronger variant of the PC$_\rho$ conjecture implies Conjecture 3 exactly, without the small polynomial factors.

Theorem 62 (PC$_\rho$ Conjecture and Conjecture 3) *Suppose that m and n are polynomial in one another and let $\epsilon > 0$ be an arbitrarily small constant. Let ρ be any one of the following distributions:*

1. *$\rho_{k\text{-PC}}(E, n, k)$ where $k = O(n^{1/2-\epsilon})$;*

2. $\rho_{\text{BPC}}(m, n, k_m, k_n)$ *where* $k_n = O(n^{1/2-\epsilon})$ *and* $k_m = O(m^{1/2-\epsilon})$*;*
3. $\rho_{k\text{-BPC}}(E, F, m, n, k_m, k_n)$ *where* $k_n = O(n^{1/2-\epsilon})$ *and* $k_m = O(m^{1/2-\epsilon})$*; and*
4. $\rho_{k\text{-HPC}^t}(E, n, k, 1/2)$ *for* $t \geq 3$ *where* $k = O(n^{1/2-\epsilon})$*.*

Then there is a constant $\delta > 0$ *such that: for any parameter* $d = O_n((\log n)^{1+\delta})$*, there is some* $p_0 = o_n(1)$ *such that* $p_\rho(s)$ *satisfies the tail bounds*

$$p_\rho(s) \leq p_0 \cdot \begin{cases} 2^{-s^2} & \text{if } 1 \leq s^2 < d \\ s^{-2d-4} & \text{if } s^2 \geq d \end{cases}$$

Proof We first prove the desired tail bounds hold for (1). Let $C > 0$ be a constant such that $k \leq Cn^{1/2-\epsilon}$. Note that the probability that S and S' independently sampled from $\rho = \rho_{k\text{-PC}}(E, n, k)$ intersect in their elements in E_i is $1/|E_i| = k/n$ for each $1 \leq i \leq k$. Furthermore, these events are independent. Thus it follows that if $\rho = \rho_{k\text{-PC}}(E, n, k)$, then p_ρ is the PMF of $\text{Bin}(k, k/n)$. In particular, we have that

$$p_\rho(s) = \binom{k}{s}\left(\frac{k}{n}\right)^s\left(1-\frac{k}{n}\right)^{k-s} \leq k^s \cdot \left(\frac{k}{n}\right)^s = \left(\frac{k^2}{n}\right)^s \leq C^{2s} \cdot n^{-2\epsilon s}$$

Let $p_0 = p_0(n)$ be a function tending to zero arbitrarily slowly. The bound above implies that $p_\rho(s) \leq p_0 \cdot 2^{-s^2}$ as long as $s \leq C_1 \log n$ for some sufficiently small constant $C_1 > 0$. Furthermore a direct computation verifies that $p_\rho(s) \leq p_0 \cdot s^{-2d-4}$ as long as

$$s \geq \frac{C_2 d \log d}{\log n}$$

for some sufficiently large constant $C_2 > 0$. Thus if $d = O_n((\log n)^{1+\delta})$ for some $\delta \in (0, 1)$, then $\frac{C_2 d \log d}{\log n} < \sqrt{d}$ and $C_1 \log n > \sqrt{d}$ for sufficiently large n. This implies the desired tail bound for (1).

The other three cases are similar. In the case of (3), if S and S' are independently sampled from $\rho = \rho_{k\text{-BPC}}(E, F, m, n, k_m, k_n)$, then the probability that S and S' intersect in their elements in E_i is k_n/n for each $1 \leq i \leq k_n$, and the probability that they intersect in their elements in $n + F_i$ is k_m/m for each $1 \leq i \leq k_m$. Thus p_ρ is distributed as independent sum of samples from $\text{Bin}(k_m, k_m/m)$ and $\text{Bin}(k_n, k_n/n)$. It follows that

$$\begin{aligned} p_\rho(s) &= \sum_{\ell=0}^{s} \binom{k_n}{\ell}\left(\frac{k_n}{n}\right)^\ell\left(1-\frac{k_n}{n}\right)^{k_n-\ell} \cdot \binom{k_m}{s-\ell}\left(\frac{k_m}{m}\right)^{s-\ell}\left(1-\frac{k_m}{m}\right)^{k_m-s+\ell} \\ &\leq \sum_{\ell=0}^{s}\left(\frac{k_n^2}{n}\right)^\ell\left(\frac{k_m^2}{m}\right)^{s-\ell} \leq s \cdot \max\left\{\left(\frac{k_n^2}{n}\right)^s, \left(\frac{k_m^2}{m}\right)^s\right\} \end{aligned} \tag{9}$$

Repeating the bounding argument as in (1) shows that the desired tail bound holds for (3) if $d = O_n((\log n)^{1+\delta})$ for some $\delta \in (0, 1)$. Since m and n are polynomial in one another implies that $\log m = \Theta(\log n)$, the $(k_m^2/m)^s$ term and the additional factor of s do not affect this bounding argument other than changing the constants C_1 and C_2. In the case of (2), similar reasoning as in

(3) yields that the distribution p_ρ where $\rho = \rho_{\text{BPC}}(m, n, k_m, k_n)$ is the independent sum of samples from Hyp (n, k_n, k_n) and Hyp (m, k_m, k_m). Now note that

$$\mathbb{P}\left[\text{Hyp}\,(n, k_n, k_n) = \ell\right] = \frac{\binom{k_n}{\ell}\binom{n-k_n}{k_n-\ell}}{\binom{n}{k_n}} \leq \frac{k_n^\ell \binom{n-\ell}{k_n-\ell}}{\binom{n}{k_n}} = k_n^\ell \prod_{i=0}^{\ell-1} \frac{k_n - i}{n-i} \leq \left(\frac{k_n^2}{n}\right)^\ell$$

This implies the same upper bound on $p_\rho(s)$ as in Equation (9) also holds for ρ in the case of (2). The argument above for (3) now establishes the desired tail bounds for (2).

We first handle the case in (4) where t is even with $t = 2r$. We have that $\rho = \rho_{k\text{-HPC}^t}(E, n, k, 1/2)$ can be sampled as $P_r^n(S) \subseteq [n^r]$ where $S \sim \mathcal{U}_n(E)$. Thus $p_\rho(s)$ is the PMF of $|P_r^n(S) \cap P_r^n(S')|$ where $S, S' \sim_{\text{i.i.d.}} \mathcal{U}_n(E)$. Furthermore the definition of P_r^n implies that $|P_r^n(S) \cap P_r^n(S')| = |S \cap S'|^r$ and, from case (1), we have that $|S \cap S'| \sim \text{Bin}(k, k/n)$. It now follows that

$$p_\rho(s) = \begin{cases} \binom{k}{s^{1/r}} \left(\frac{k}{n}\right)^{s^{1/r}} \left(1 - \frac{k}{n}\right)^{k - s^{1/r}} & \text{if } s \text{ is an } r\text{th power} \\ 0 & \text{otherwise} \end{cases}$$

The same bounds as in case (1) therefore imply that $p_\rho(s) \leq (k^2/n)^{s^{1/r}}$ for all $s \geq 0$. A similar analysis as in (1) now shows that $p_\rho(s) \leq p_0 \cdot 2^{-s^2}$ holds if $s \leq C_1 (\log n)^{r/(2r-1)}$ for some sufficiently small constant $C_1 > 0$, and that $p_\rho(s) \leq p_0 \cdot s^{-2d-4}$ holds if

$$s \geq C_2 \left(\frac{d \log d}{\log n}\right)^r$$

for some sufficiently large constant $C_2 > 0$. As long as $d = O_n((\log n)^{1+\delta})$ for some $0 < \delta < 1/(2r-1)$, we have that $C_2 \left(\frac{d \log d}{\log n}\right)^r < \sqrt{d}$ and $C_1 (\log n)^{r/(2r-1)} > \sqrt{d}$ for sufficiently large n. Since t and r are constants here, δ can be taken to be constant as well. In the case where t is odd, it follows that $\rho_{k\text{-HPC}^t}(E, n, k, 1/2)$ is the same as $\rho_{k\text{-HPC}^{2t}}(F, \sqrt{n}, \sqrt{k}, 1/2)$ for some partition F as long as n and k are squares. The same argument establishes the desired tail bound for this prior, completing the case of (4) and proof of the theorem. ■

K.2. Low-Degree Polynomials and the PC$_\rho$ Conjecture

In this section, we show that the low-degree conjecture – that low-degree polynomials are optimal for a class of average-case hypothesis testing problems – implies the PC$_\rho$ conjecture. In particular, we will obtain a simple expression capturing the power of the optimal low-degree polynomial for PC$_\rho$ in Proposition 67. We then will apply this proposition to prove Theorem 68, showing that the power of this optimal low-degree polynomial tends to zero under the tail bounds on p_ρ in the PC$_\rho$ conjecture. We also will discuss a stronger version of the PC$_\rho$ conjecture that exactly implies Conjecture 3. First, we informally introduce the low-degree conjecture and the technical conditions arising in its various formalizations in the literature.

Polynomial Tests and the Low-Degree Conjecture. In this section, will draw heavily from similar discussions in Hopkins and Steurer (2017) and Hopkins's thesis Hopkins (2018). Throughout, we will consider discrete hypothesis testing problems with observations taken without loss of generality to lie in the discrete hypercube $\{-1, 1\}^N$. For example, an n-vertex instance of planted clique

can be represented in the discrete hypercube by the above-diagonal entries of its signed adjacency matrix when $N = \binom{n}{2}$. Given a hypothesis H_0, the term D-simple statistic refers to polynomials $f : \{-1,1\}^N \to \mathbb{R}$ of degree at most D in the coordinates of $\{-1,1\}^N$ that are calibrated and normalized so that $\mathbb{E}_{H_0} f(X) = 0$ and $\mathbb{E}_{H_0} f(X)^2 = 1$.

For a broad range of hypothesis testing problems, it has been observed in the literature that D-simple statistics seem to capture the full power of the SOS hierarchy (Hopkins and Steurer, 2017; Hopkins, 2018). This trend prompted a further conjecture that D-simple statistics often capture the full power of efficient algorithms, leading more concretely to the *low-degree conjecture* which is stated informally below. This conjecture has been used to gather evidence of hardness for a number of natural detection problems and has generally emerged as a convenient tool to predict statistical-computational gaps (Hopkins and Steurer, 2017; Hopkins, 2018; Kunisky et al., 2019; Bandeira et al., 2019). Variants of this low-degree conjecture have appeared as Hypothesis 2.1.5 and Conjecture 2.2.4 in Hopkins (2018) and Conjectures 1.16 and 4.6 in Kunisky et al. (2019).

Conjecture 63 (Informal – Hypothesis 2.1.5 in Hopkins (2018)) *For a broad class of hypothesis testing problems H_0 versus H_1, there is a test running in time $N^{\tilde{O}(D)}$ with Type I+II error tending to zero if and only if there is a successful D-simple statistic i.e. a polynomial f of degree at most D such that $\mathbb{E}_{H_0} f(X) = 0$ and $\mathbb{E}_{H_0} f(X)^2 = 1$ yet $\mathbb{E}_{H_1} f(X) \to \infty$.*

Detailed discussions of the low-degree conjecture and the connections between D-simple statistics and other types of algorithms can be found in Kunisky et al. (2019) and Holmgren and Wein (2020). The informality in the conjecture above is the undefined "broad class" of hypothesis testing problems. In Hopkins (2018), several candidate technical conditions defining this class were proposed and subsequently have been further refined in Kunisky et al. (2019) and Holmgren and Wein (2020). These conditions are discussed in more detail later in this section.

The utility of the low-degree conjecture in predicting statistical-computational gaps arises from the fact that the optimal D-simple statistic can be explicitly characterized. By the Neyman-Pearson lemma, the optimal test with respect to Type I+II error is the the likelihood ratio test, which declares H_1 if $\mathsf{LR}(X) = \mathbb{P}_{H_1}(X)/\mathbb{P}_{H_0}(X) > 1$ and H_0 otherwise, given a sample X. Computing the likelihood ratio is typically intractable in problems in high-dimensional statistical inference. The low-degree likelihood ratio $\mathsf{LR}^{\leq D}$ is the orthogonal projection of the likelihood ratio onto the subspace of polynomials of degree at most D. When H_0 is a product distribution on the discrete hypercube $\{-1,1\}^N$, the following theorem asserts that $\mathsf{LR}^{\leq D}$ is the optimal test of a given degree. Here, the projection is with respect to the inner product $\langle f, g\rangle = \mathbb{E}_{H_0} f(X) g(X)$, which also defines a norm $\|f\|_2^2 = \langle f, f\rangle$.

Theorem 64 (Page 35 of Hopkins (2018)) *The optimal D-simple statistic is the low-degree likelihood ratio, i.e. it holds that*

$$\max_{\substack{f \in \mathbb{R}[x]_{\leq D} \\ \mathbb{E}_{H_0} f(X) = 0}} \frac{\mathbb{E}_{H_1} f(X)}{\sqrt{\mathbb{E}_{H_0} f(X)^2}} = \|\mathsf{LR}^{\leq D} - 1\|_2$$

Thus existence of low-degree tests for a given problem boils down to computing the norm of the low-degree likelihood ratio. When H_0 is the uniform distribution on $\{-1,1\}^N$, the norm above can be re-expressed in terms of the standard Boolean Fourier basis. Let the collection of functions

$\{\chi_\alpha(X) = \prod_{e\in\alpha} X_e : \alpha \subseteq [N]\}$ denote this basis, which is orthonormal over the space $\{-1,1\}^N$ with inner product defined above. By orthonormality, any χ_α with $1 \le |\alpha| \le D$ satisfies that

$$\langle \chi_\alpha, \mathsf{LR}^{\le D} - 1\rangle = \langle \chi_\alpha, \mathsf{LR}\rangle = \mathbb{E}_{H_0}\chi_\alpha(X)\mathsf{LR}(X) = \mathbb{E}_{H_1}\chi_\alpha(X)$$

and $\mathbb{E}_{H_0}\mathsf{LR}^{\le D} = \mathbb{E}_{H_1}1 = 1$ so that $\langle 1, \mathsf{LR}^{\le D} - 1\rangle = 0$. It then follows by Parseval's identity that

$$\|\mathsf{LR}^{\le D} - 1\|_2 = \left(\sum_{1\le|\alpha|\le D} \left(\mathbb{E}_{H_1}\chi_\alpha(X)\right)^2\right)^{1/2} \tag{10}$$

which is exactly the Fourier energy up to degree D.

Technical Conditions, S_n-Invariance and Counterexamples. While Conjecture 63 is believed to accurately predict the computational barriers in nearly any natural high-dimensional statistical problem including all of the problems we consider, a precise set of criteria exactly characterizing this "broad class" has yet to be pinned down in the literature. The following was the first formalization of the low-degree conjecture, which appeared as Conjecture 2.2.4 in Hopkins (2018).

Conjecture 65 (Conjecture 2.2.4 in Hopkins (2018)) *Let Ω be a finite set or $\mathbb{R}$, and let k be a fixed integer. Let $N = \binom{n}{k}$, let ν be a product distribution on Ω^N and let μ be another distribution on Ω^N. Suppose that μ is S_n-invariant and $(\log n)^{1+\Omega(1)}$-wise almost independent with respect to ν. Then no polynomial time test distinguishes $T_\delta\mu$ and ν with probability $1 - o(1)$, for any $\delta > 0$. Formally, for all $\delta > 0$ and every polynomial-time test $t : \Omega^N \to \{0,1\}$ there exists $\delta' > 0$ such that for every large enough n,*

$$\frac{1}{2}\mathbb{P}_{x\sim\nu}\left[t(x) = 0\right] + \frac{1}{2}\mathbb{P}_{x\sim T_\delta\mu}\left[t(x) = 1\right] \le 1 - \delta'$$

This conjecture has several key technical stipulations attempting to conservatively pin down the $\tilde{O}$ in Conjecture 63 and a set of *sufficient conditions* to be in this "broad class". We highlight and explain these key conditions below.

1. The distribution μ is required to be S_n-invariant. Here, a distribution μ on Ω^N is said to be S_n-invariant if $\mathbb{P}_\mu(x) = \mathbb{P}_\mu(\pi \cdot x)$ for all $\pi \in S_n$ and $x \in \Omega^N$, where π acts on x by identifying the coordinates of x with the k-subsets of $[n]$ and permuting these coordinates according to the permutation on k-subsets induced by π.

2. The $(\log n)^{1+\Omega(1)}$-wise almost independence requirement on μ essentially enforces that polynomials of degree at most $(\log n)^{1+\Omega(1)}$ are unable to distinguish between μ and ν. More formally, a distribution μ is D-wise almost independent with respect to ν if every D-simple statistic f, calibrated and normalized with respect to ν, satisfies that $\mathbb{E}_{x\sim\mu}f(x) = O(1)$.

3. Rather than μ, the distribution the conjecture asserts is hard to distinguish from ν is the result $T_\delta\mu$ of applying the noise operator T_δ. Here, the distribution $T_\delta\mu$ is defined by first sampling $x \sim \mu$, then sampling $y \sim \nu$ and replacing each x_i with y_i independently with probability δ.

These technical conditions are intended to conservatively rule out specific pathological examples. As mentioned in Hopkins (2018), the purpose of T_δ is to destroy algebraic structure that may lead

to efficient algorithms that cannot be implemented with low-degree polynomials. For example, if μ uniform over the solution set to a satisfiable system of equations mod 2 and ν is the uniform distribution, it is possible to distinguish these two distributions through Gaussian elimination while the lowest D for which a D-simple statistic does so can be as large as $D = \Omega(N)$. The noise operator T_δ rules out distributions with this kind of algebraic structure. The $(\log n)^{1+\Omega(1)}$-wise requirement on the almost independence of μ and the $\tilde{O}(D)$ in Conjecture 63 are both to account for the fact that some common polynomial time algorithms for natural hypothesis testing problems can only be implemented as degree $O(\log n)$ polynomials. For example, Section 4.2.3 of Kunisky et al. (2019) shows that spectral methods can typically be implemented as degree $O(\log n)$ polynomials.

In Hopkins (2018), it was mentioned that the S_n-invariance condition was included in Conjecture 65 mainly because most canonical inference problems satisfy this property and, furthermore, that there were no existing counterexamples to the conjecture without it. Recently, Holmgren and Wein (2020) gave two construction of hypothesis testing problems based on efficiently-correctable binary codes and Reed-Solomon codes. The first construction is for binary Ω and admits a polynomial-time test despite being $\Omega(n)$-wise almost independent. This shows that T_δ is insufficient to always rule out high-degree algebraic structure that can be used in efficient algorithms. However, this construction also is highly asymmetric and ruled out by S_n-invariance condition in Conjecture 65. The second construction is for $\Omega = \mathbb{R}$ and admits a polynomial-time test despite being both $\Omega(n)$-wise almost independent and S_n-invariant, thus falsifying Conjecture 65 as stated. However, as discussed in Holmgren and Wein (2020), the conjecture can easily be remedied by replacing T_δ with another operator, such as the Ornstein-Uhlenbeck noise operator. In this work, only the case of binary Ω will be relevant to the PC_ρ conjecture.

The PC_ρ Conjecture, Technical Conditions and a Generalization. The PC_ρ hypothesis testing problems and their planted dense subgraph generalizations PDS_ρ that we consider in this work can be shown to satisfy a wide range of properties sufficient to rule out known counterexamples to the low-degree conjecture. In particular, these problems almost satisfy all three conservative conditions proposed in Hopkins (2018), instead satisfying a milder requirement for sufficient symmetry than full S_n-invariance.

1. By definition, a general instance of PC_ρ with an arbitrary ρ is only invariant to permutations $\pi \in S_n$ that ρ is also invariant to. However, each of the specific hardness assumptions we use in our reductions corresponds to a ρ with a large amount of symmetry and that is invariant to large subgroups of S_n. For example, k-PC and k-PDS are invariant to permutations within each part E_i, each of which has size $n/k = \omega(\sqrt{n})$. This symmetry seems sufficient to break the error-correcting code approach used to construct counterexamples to the low-degree conjecture in Holmgren and Wein (2020).

2. As will be shown subsequently in this section, the conditions in the PC_ρ conjecture imply that a PC_ρ instance be $(\log n)^{1+\Omega(1)}$-wise almost independent for it to be conjectured to be hard.

3. While PC_ρ is not of the form $T_\delta \mu$, its generalization PDS_ρ at any pair of constant edge densities $0 < q < p < 1$ always is. All of our reductions also apply to input instances of PDS_ρ and thus a PDS_ρ variant of the PC_ρ conjecture is sufficient to deduce our computational lower bounds. That said, we do not expect that the computational complexity of PC_ρ and PDS_ρ to be different as long as p and q are constant.

As mentioned in Section 2, while we restrict our formal statement of the PC_ρ conjecture to the specific hardness assumptions we need for our reductions, we believe it should hold generally for ρ with sufficient symmetry. A candidate condition is that ρ is invariant to a subgroup $H \subseteq S_n$ of permutations such that, for each index $i \in [n]$, there are at least $n^{\Omega(n)}$ permutations $\pi \in H$ with $\pi(i) \neq i$. This ensures that ρ has a large number of nontrivial symmetries that are not just permuting coordinates known not to lie in the clique.

We also remark that there are many examples of hypothesis testing problems where the three conditions in Hopkins (2018) are violated but low-degree polynomials still seem to accurately predict the performance of the best known efficient algorithms. As mentioned in Holmgren and Wein (2020), the spiked Wishart model does not quite satisfy S_n-invariance but still low-degree predictions are conjecturally accurate. Ordinary PC is not of the form $T_\delta \mu$ and the low-degree conjecture accurately predicts the PC conjecture, which is widely believed to be true.

The Degree Requirement and a Stronger PC_ρ Conjecture. Furthermore, the degree requirement for the almost independence condition of Conjecture 65 is often not exactly necessary. It is discussed in Section 4.2.5 of Kunisky et al. (2019) that, for sufficiently nice distributions H_0 and H_1, low-degree predictions are often still accurate when the almost independence condition is relaxed to only be $\omega(1)$-wise for any $\omega(1)$ function of n. This yields the following stronger variant of the PC_ρ conjecture.

Conjecture 66 (Informal – Stronger PC_ρ Conjecture) *For sufficiently symmetric ρ, there is no polynomial time algorithm solving $\text{PC}_\rho(n, k, 1/2)$ if there is some function $w(n) = \omega_n(1)$ such that the tail bounds on $p_\rho(s)$ in Conjecture 2 are only guaranteed to hold for all $d \leq w(n)$.*

We conjecture that the ρ in Conjecture 3 are symmetric enough for this conjecture to hold. A nearly identical argument to that in Theorem 62 can be used to show that this stronger PC_ρ conjecture implies the exact boundaries in Conjecture 3, without the small polynomial error factors of $O(n^\epsilon)$ and $O(m^\epsilon)$.

We now make several notes on the degree requirement in the PC_ρ conjecture, as stated in Conjecture 2. As will be shown later in this section, the tail bounds on $p_\rho(s)$ for a particular d directly imply the d-wise almost independence of PC_ρ. Now note that for any ρ and $k \gg \log n$, there is always a d-simple statistic solving PC_ρ with $d = O((\log n)^2)$. Specifically, $\mathcal{G}(n, 1/2)$ has its largest clique of size less than $(2 + \epsilon) \log_2 n$ with probability $1 - o_n(1)$ and any instance of H_1 of PC_ρ with $k \gg \log n$ has $n^{\omega(1)}$ cliques of size $\lceil 3 \log_2 n \rceil$. Furthermore, the number of cliques of this size can be expressed as a degree $O((\log n)^2)$ polynomial in the edge indicators of a graph. Similarly, the largest clique in an s-uniform Erdős-Rényi hypergraph is in general of size $O((\log n)^{1/(s-1)})$ and a simple clique-counting test distinguishing this from the planted clique hypergraph distribution can be expressed as an $O((\log n)^{s/(s-1)})$ degree polynomial. This shows that for all ρ, the problem PC_ρ is not $O((\log n)^2)$-wise almost independent. Furthermore, for any $\delta > 0$, there is some ρ corresponding to a hypergraph variant of PC such that PC_ρ is not $O((\log n)^{1+\delta})$-wise almost independent. Thus the tail bounds in Conjecture 2 never hold for $\delta \geq 1$ and, for any $\delta' > 0$, there is some ρ requiring $\delta \leq \delta'$ for these tail bounds to be true.

Finally, we remark that there are highly asymmetric examples of ρ for which Conjecture 66 is not true. Suppose that n is even, let $c > 0$ be an arbitrarily large integer and let $S_1, S_2, \ldots, S_{n^c} \subseteq [n/2]$ be a known family of subsets of size $\lceil 3 \log_2 n \rceil$. Now let ρ be sampled by taking the union of an S_i chosen uniformly at random and a size $k - \lceil 3 \log_2 n \rceil$ subset of $\{n/2 + 1, n/2 + 2, \ldots, n\}$

chosen uniformly at random. The resulting PC_ρ problem can be solved in polynomial time by exhaustively searching for the subset S_i. However, this ρ only violates the tail bounds on p_ρ in Conjecture 2 for $d = \Omega_n(\log n/\log\log n)$. If $S_1, S_2, \ldots, S_{n^c}$ are sufficiently pseudorandom, then the structure of this ρ only appears in the tails of $p_\rho(s)$ when $s \geq \lceil 3\log_2 n\rceil$. In particular, the probability that $s \geq \lceil 3\log_2 n\rceil$ under p_ρ is at least the chance that two independent samples from ρ choose the same S_i, which occurs with probability n^{-c}. It can be verified the the tail bound of $p_0 \cdot s^{-2d-4}$ in Conjecture 2 only excludes this possibility when $d = \Omega_n(\log n/\log\log n)$. We remark though that this ρ is highly asymmetric and any mild symmetry assumption that would effectively cause the number of S_i to be super-polynomial would break this example.

The Low-Degree Conjecture and PC_ρ. We now will characterize the power of the optimal D-simple statistics for PC_ρ. The following proposition establishes an explicit formula for $\mathsf{LR}^{\leq D}$ in PC_ρ, which will be shown in the subsequent theorem to naturally yield the PMF decay condition in the PC_ρ conjecture.

Proposition 67 *Let $\mathsf{LR}^{\leq D}$ be the low-degree likelihood ratio for the hypothesis testing problem $\text{PC}_\rho(n,k,1/2)$ between $\mathcal{G}(n,1/2)$ and $\mathcal{G}_\rho(n,k,1/2)$. For any $D \geq 1$, it follows that*

$$\|\mathsf{LR}^{\leq D} - 1\|_2^2 = \mathbb{E}_{S,S'\sim\rho^{\otimes 2}}\left[\#\text{ of nonempty edge subsets of } S\cap S' \text{ of size at most } D\right]$$

Proof In the notation above, let $N = \binom{n}{2}$ and identify $X \in \{-1,1\}^N$ with the space of signed adjacency matrices X of n-vertex graphs. Let P_S be the distribution on graphs in this space induced by $\text{PC}(n,k,1/2)$ conditioned on the clique being planted on the vertices in the subset S i.e. such that $X_{ij} = 1$ if $i \in S$ and $j \in S$ and otherwise $X_{ij} = \pm 1$ with probability half each. Now let $\alpha \subseteq \mathcal{E}_0$ be a subset of possible edges. The set of functions $\{\chi_\alpha(X) = \prod_{e\in\alpha} X_e : \alpha \subseteq \mathcal{E}_0\}$ comprises the standard Fourier basis on $\{-1,1\}^{\mathcal{E}_0}$. For each fixed clique S, because $\mathbb{E}_{P_S} X_e = 0$ if $e \notin \binom{S}{2}$ and non-clique edges are independent, we see that

$$\mathbb{E}_{P_S}[\chi_\alpha(X)] = \mathbf{1}\{V(\alpha) \subseteq S\}$$

We therefore have that

$$\mathbb{E}_{H_1}[\chi_\alpha(X)] = \mathbb{E}_{S\sim\rho}\mathbb{E}_{P_S}[\chi_\alpha(X)] = \mathbb{E}_{S\sim\rho}\left[\mathbf{1}\{V(\alpha)\subseteq S\}\right] = \mathbb{P}_\rho\left[V(\alpha)\subseteq S\right]$$

Now suppose that S' is drawn from ρ independently of S. It now follows that

$$\begin{aligned}
\mathbb{E}_{H_1}[\chi_\alpha(X)]^2 &= \mathbb{E}_{S\sim\rho}\left[\mathbf{1}\{V(\alpha)\subseteq S\}\right]^2 \\
&= \mathbb{E}_{S\sim\rho}\left[\mathbf{1}\{V(\alpha)\subseteq S\}\right]\cdot\mathbb{E}_{S'\sim\rho}\left[\mathbf{1}\{V(\alpha)\subseteq S'\}\right] \\
&= \mathbb{E}_{S,S'\sim\rho^{\otimes 2}}\left[\mathbf{1}\{V(\alpha)\subseteq S\}\cdot\mathbf{1}\{V(\alpha)\subseteq S'\}\right] \\
&= \mathbb{E}_{S,S'\sim\rho^{\otimes 2}}\left[\mathbf{1}\left\{V(\alpha)\subseteq S\cap S'\right\}\right]
\end{aligned}$$

From Equation (10), we therefore have that

$$\|\mathsf{LR}^{\leq D} - 1\|_2^2 = \sum_{1\leq|\alpha|\leq D}\mathbb{E}_{H_1}\left[\chi_\alpha(X)\right]^2 = \mathbb{E}_{S,S'\sim\rho^{\otimes 2}}\left[\sum_{1\leq|\alpha|\leq D}\mathbf{1}\left\{V(\alpha)\subseteq S\cap S'\right\}\right]$$

Now observe that the sum

$$\sum_{1\le|\alpha|\le D} \mathbf{1}\left\{V(\alpha)\subseteq S\cap S'\right\}$$

counting the number of nonempty edge subsets of $S\cap S'$ of size at most D. ■

This proposition now allows us to show the main result of this section, which is that the condition in the PC_ρ conjecture is enough to show the failure of low-degree polynomials for PC_ρ. Combining the next theorem with Conjecture 63 would suggest that whenever the PMF decay condition of the PC_ρ condition holds, there is no polynomial time algorithm solving $\mathrm{PC}_\rho(n,k,1/2)$.

Theorem 68 (PC_ρ Implies Failure of Low-Degree) *Suppose that ρ satisfies that for any parameter $d=O_n(\log n)$, there is some $p_0=o_n(1)$ such that $p_\rho(s)$ satisfies the tail bounds*

$$p_\rho(s)\le p_0\cdot\begin{cases}2^{-s^2} & \text{if } 1\le s^2<d\\ s^{-2d-4} & \text{if } s^2\ge d\end{cases}$$

Let $\mathsf{LR}^{\le D}$ be the low-degree likelihood ratio for the hypothesis testing problem $\mathrm{PC}_\rho(n,k,1/2)$. Then it also follows that for any parameter $D=O_n(\log n)$, we have

$$\|\mathsf{LR}^{\le D}-1\|_2=o_n(1)$$

Proof First observe that the number of nonempty edge subsets of $S\cap S'$ of size at most D can be expressed explicitly as

$$f_D(s)=\sum_{\ell=1}^{D}\binom{s(s-1)/2}{\ell}$$

if $s=|S\cap S'|$. Furthermore, we can crudely upper bound f_D in two separate ways. Note that the number of nonempty edge subsets of $S\cap S'$ is exactly $2^{\binom{s}{2}}-1$ if $s=|S\cap S'|$. Therefore we have that $f_D(s)\le 2^{\binom{s}{2}}$. Furthermore using the upper bound that $\binom{x}{\ell}\le x^\ell$, we have that if $s\ge 3$ then

$$f_D(s)=\sum_{\ell=1}^{D}\binom{s(s-1)/2}{\ell}\le\sum_{\ell=1}^{D}\left(\frac{s(s-1)}{2}\right)^\ell\le\frac{\left(\frac{s(s-1)}{2}\right)^{D+1}-1}{\left(\frac{s(s-1)}{2}\right)-1}\le s^{2(D+1)}$$

Combining these two crude upper bounds, we have that $f_D(s)\le\min\left\{2^{\binom{s}{2}},s^{2(D+1)}\right\}$. Also note that $f_D(0)=f_D(1)=0$. Combining this with the given bounds on $p_\rho(s)$, we have that

$$\begin{aligned}\|\mathsf{LR}^{\le D}-1\|_2^2&=\mathbb{E}_{S,S'\sim\rho^{\otimes 2}}\left[f_D(|S\cap S'|)\right]\\&=\sum_{s=2}^{k}p_\rho(s)\cdot f_D(s)\\&\le p_0\cdot\sum_{1\le s^2<D}2^{-s^2}\cdot f_D(s)+p_0\cdot\sum_{D\le s^2\le k^2}s^{-2d-4}\cdot f_D(s)\\&\le p_0\cdot\sum_{1\le s^2<D}2^{-s^2}\cdot 2^{\binom{s}{2}}+p_0\cdot\sum_{D\le s^2\le k^2}s^{-2D-4}\cdot s^{2(D+1)}\end{aligned}$$

$$= p_0 \cdot \sum_{s=1}^{\infty} 2^{-\binom{s+1}{2}} + p_0 \cdot \sum_{s=1}^{\infty} s^{-2} = O_n(p_0)$$

which completes the proof of the theorem. ■

K.3. Statistical Query Algorithms and the PC$_\rho$ Conjecture

In this section, we verify that the lower bounds shown by Feldman et al. (2013) for PC for a generalization of statistical query algorithms hold essentially unchanged for SQ variants of k-PC, k-BPC and BPC. We remark at the end of this section why the statistical query model seems ill-suited to characterizing the computational barriers in problems that are tensor or hypergraph problems such as k-HPC. Since it was shown in Section K.1 that there are specific ρ in PC$_\rho$ corresponding to k-HPC, it similarly follows that the SQ model seems ill-suited to characterizing the barriers PC$_\rho$ for general ρ. Throughout this section, we focus on k-PC, as lower bounds in the statistical query model for k-BPC and BPC will follow from nearly identical arguments.

Distributional Problems and SQ Dimension. The Statistical Algorithm framework of Feldman et al. (2013) applies to distributional problems, where the input is a sequence of i.i.d. observations from a distribution D. In order to obtain lower bounds in the statistical query model supporting Conjecture 3, we need to define a distributional analogue of k-PC. As in Feldman et al. (2013), a natural distributional version can be obtained by considering a bipartite version of k-PC, which we define as follows.

Definition 69 (Distributional Formulation of k-PC) *Let k divide n and fix a known partition E of $[n]$ into k parts $E_1, E_2, \ldots, E_k$ with $|E_i| = n/k$. Let $S \subseteq [n]$ be a subset of indices with $|S \cap E_i| = 1$ for each $i \in [k]$. The distribution D_S over $\{0,1\}^n$ produces with probability $1 - k/n$ a uniform point $X \sim \mathrm{Unif}(\{0,1\}^n)$ and with probability k/n a point X with $X_i = 1$ for all $i \in S$ and $X_{S^c} \sim \mathrm{Unif}(\{0,1\})^{n-k}$. The distributional bipartite k-PC problem is to find the subset S given some number of independent samples m from D_S.*

In other words, the distribution k-PC problem is k-BPC with n left and n right vertices, a randomly-sized right part of the planted biclique and no k-partite structure on the right vertex set. We remark that many of our reductions, such as our reductions to RSME, NEG-SPCA, MSLR and RSLSR, only need the k-partite structure along one vertex set of k-PC or k-BPC. This distributional formulation of k-PC is thus a valid starting point for these reductions.

We now formally introduce the Statistical Algorithm framework of Feldman et al. (2013) and SQ dimension. Let $\mathcal{X} = \{0,1\}^n$ denote the space of configurations and let $\mathcal{D}$ be a set of distributions over $\mathcal{X}$. Let $\mathcal{F}$ be a set of solutions and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map taking each distribution $D \in \mathcal{D}$ to a subset of solutions $\mathcal{Z}(D) \subseteq \mathcal{F}$ that are defined to be valid solutions for D. In our setting, $\mathcal{F}$ corresponds to clique positions S respecting the partition E. Furthermore, since each clique position is in one-to-one correspondence with distributions, there is a single clique $\mathcal{Z}(D)$ corresponding to each distribution D. For $m > 0$, the *distributional search problem* $\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$ using m samples is to find a valid solution $f \in \mathcal{Z}(D)$ given access to m random samples from an unknown $D \in \mathcal{D}$.

Classes of algorithms in the framework of Feldman et al. (2013) are defined in terms of access to oracles. The most basic oracle is an unbiased oracle, which evaluates a simple function on a single sample as follows.

Definition 70 (Unbiased Oracle) *Let D be the true unknown distribution. A query to the oracle consists of any function $h : \mathcal{X} \to \{0, 1\}$, and the oracle then takes an independent random sample $X \sim D$ and returns $h(X)$.*

Algorithms with access to an unbiased oracle are referred to as *unbiased statistical algorithms*. Since these algorithms access the sampled data only through the oracle, it is possible to prove *unconditional* lower bounds using information-theoretic methods. Another oracle is the $VSTAT$, defined below, which is similar but also allowed to make an adversarial perturbation of the function evaluation. It is shown in Feldman et al. (2013) via a simulation argument that the two oracles are approximately equivalent.

Definition 71 ($VSTAT$ Oracle) *Let D be the true distribution and $t > 0$ a sample size parameter. A query to the $VSTAT(t)$ oracle consists of any function $h : \mathcal{X} \to [0, 1]$, and the oracle returns an arbitrary value $v \in [\mathbb{E}_D h(X) - \tau, \mathbb{E}_D h(X) + \tau]$, where $\tau = \max\{1/t, \sqrt{\mathbb{E}_D h(X)(1 - \mathbb{E}_D h(X))/t}\}$.*

We borrow some definitions from Feldman et al. (2013). Given a distribution D, we define the inner product $\langle f, g \rangle_D = \mathbb{E}_{X \sim D} f(X) g(X)$ and the corresponding norm $\|f\|_D = \sqrt{\langle f, f \rangle_D}$. Given two distributions D_1 and D_2 both absolutely continuous with respect to D, their pairwise correlation is defined to be

$$\chi_D(D_1, D_2) = \left| \left\langle \frac{D_1}{D} - 1, \frac{D_2}{D} - 1 \right\rangle_D \right| = \left| \langle \widehat{D}_1, \widehat{D}_2 \rangle_D \right| .$$

where $\widehat{D}_1 = \frac{D_1}{D} - 1$. The *average correlation* $\rho(\mathcal{D}, D)$ of a set of distributions $\mathcal{D}$ relative to distribution D is then given by

$$\rho(\mathcal{D}, D) = \frac{1}{|\mathcal{D}|^2} \sum_{D_1, D_2 \in \mathcal{D}} \chi_D(D_1, D_2) = \frac{1}{|\mathcal{D}|^2} \sum_{D_1, D_2 \in \mathcal{D}} \left| \left\langle \frac{D_1}{D} - 1, \frac{D_2}{D} - 1 \right\rangle_D \right| .$$

Given these definitions, we can now introduce the key quantity from Feldman et al. (2013), statistical dimension, which is defined in terms of average correlation.

Definition 72 (Statistical dimension) *Fix $\gamma > 0, \eta > 0$, and search problem $\mathcal{Z}$ over set of solutions $\mathcal{F}$ and class of distributions $\mathcal{D}$ over $\mathcal{X}$. We consider pairs $(D, \mathcal{D}_D)$ consisting of a "reference distribution" D over $\mathcal{X}$ and a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ with the following property: for any solution $f \in \mathcal{F}$, the set $\mathcal{D}_f = \mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)$ has size at least $(1 - \eta) \cdot |\mathcal{D}_D|$. Let $\ell(D, \mathcal{D}_D)$ be the largest integer ℓ so that for any subset $\mathcal{D}' \subseteq \mathcal{D}_f$ with $|\mathcal{D}'| \geq |\mathcal{D}_f|/\ell$, the average correlation is $|\rho(\mathcal{D}', D)| < \gamma$ (if there is no such ℓ one can take $\ell = 0$). The* statistical dimension *with average correlation γ and solution set bound η is defined to be the largest $\ell(D, \mathcal{D}_D)$ for valid pairs $(D, \mathcal{D}_D)$ as described, and is denoted by* $\mathrm{SDA}(\mathcal{Z}, \gamma, \eta)$.

In Feldman et al. (2013), it is shown that statistical dimension immediately yields a lower bound on the number of queries to an unbiased oracle or a $VSTAT$ oracle needed to solve a given distributional search problem.

Theorem 73 (Theorems 2.7 and 3.17 of Feldman et al. (2013)) *Let $\mathcal{X}$ be a domain and $\mathcal{Z}$ a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over $\mathcal{X}$. For $\gamma > 0$ and $\eta \in (0, 1)$, let $\ell = \mathrm{SDA}(\mathcal{Z}, \gamma, \eta)$. Any (possibly randomized) statistical query algorithm that solves $\mathcal{Z}$ with probability $\delta > \eta$ requires at least ℓ calls to the $VSTAT(1/(3\gamma))$ oracle to solve $\mathcal{Z}$.*

Moreover, any statistical query algorithm requires at least m calls to the Unbiased Oracle for $m = \min\left\{\frac{\ell(\delta-\eta)}{2(1-\eta)}, \frac{(\delta-\eta)^2}{12\gamma}\right\}$. In particular, if $\eta \le 1/6$, then any algorithm with success probability at least $2/3$ requires at least $\min\{\ell/4, 1/48\gamma\}$ samples from the Unbiased Oracle.

We remark that the number of queries to an oracle is a lower bound on the runtime of the statistical algorithm in question. Furthermore, the number of "samples" m corresponding to a $VSTAT(t)$ oracle is t, as this is the number needed to approximately obtain the confidence interval of width 2τ in the definition of the $VSTAT$ oracle above.

SQ Lower Bounds for Distributional k-PC. We now will use the theorem above to deduce SQ lower bounds for distributional k-PC. Let $\mathcal{S}$ be the set of all k-subsets of $[n]$ respecting the partition E i.e. $\mathcal{S} = \{S : |S| = k \text{ and } |S \cap E_i| = 1 \text{ for } i \in [k]\}$. Note that $|\mathcal{S}| = (n/k)^k$. We henceforth use D to denote the uniform distribution on $\{0,1\}^n$. The following lemma is as in Feldman et al. (2013), except that we further restrict S and T to be in $\mathcal{S}$ rather than arbitrary size k subsets of $[n]$, which does not change the bound.

Lemma 74 (Lemma 5.1 in Feldman et al. (2013)) *For $S, T \in \mathcal{S}$, $\chi_D(D_S, D_T) = |\langle \widehat{D}_S, \widehat{D}_T \rangle_D| \le 2^{|S\cap T|}k^2/n^2$.*

The following lemma is crucial to deriving the SQ dimension of distributional k-PC and is similar to Lemma 5.2 in Feldman et al. (2013). Its proof is deferred to Appendix R.1.

Lemma 75 (Modification of Lemma 5.2 in Feldman et al. (2013)) *Let $\delta \ge 1/\log n$ and $k \le n^{1/2-\delta}$. For any integer $\ell \le k$, $S \in \mathcal{S}$, and set $A \subseteq \mathcal{S}$ with $|A| \ge 2|\mathcal{S}|/n^{2\ell\delta}$,*

$$\frac{1}{|A|}\sum_{T\in A}\left|\langle \widehat{D}_S, \widehat{D}_T\rangle_D\right| \le 2^{\ell+3}\frac{k^2}{n^2}\,.$$

This lemma now implies the following SQ dimension lower bound for distributional k-PC.

Theorem 76 (Analogue of Theorem 5.3 of Feldman et al. (2013)) *For $\delta \ge 1/\log n$ and $k \le n^{1/2-\delta}$, let $\mathcal{Z}$ denote the distributional bipartite k-PC problem. If $\ell \le k$, then it follows that $SDA(\mathcal{Z}, 2^{\ell+3}k^2/n^2, \left(\frac{n}{k}\right)^{-k}) \ge n^{2\ell\delta}/8$.*

Proof For each clique position S let $\mathcal{D}_S = \mathcal{D} \setminus \{D_S\}$. Then $|\mathcal{D}_S| = \left(\frac{n}{k}\right)^k - 1 = \left(1 - \left(\frac{n}{k}\right)^{-k}\right)|\mathcal{D}|$. Now for any $\mathcal{D}'$ with $|\mathcal{D}'| \ge 2|\mathcal{S}|/n^{2\ell\delta}$ we can apply Lemma 75 to conclude that $\rho(\mathcal{D}', D) \le 2^{\ell+3}k^2/n^2$. By Definition 72 of statistical dimension this implies the bound stated in the theorem. ■

Applying Theorem 73 to this statistical dimension lower bound yields the following hardness for statistical query algorithms.

Corollary 77 (SQ Lower Bound for Recovery in Distributional k-PC) *For any constant $\delta > 0$ and $k \le n^{1/2-\delta}$, any SQ algorithm that solves the distributional bipartite k-PC problem requires $\Omega(n^2/k^2 \log n) = \tilde{\Omega}(n^{1+2\delta})$ queries to the Unbiased Oracle.*

This is to be interpreted as impossible, as there are only n right vertices vertices available in the actual bipartite graph. Because all the quantities in Theorem 76 are the same as in Feldman et al. (2013) up to constants, the same logic as used there allows to deduce a statement regarding the hypothesis testing version, stated there as Theorems 2.9 and 2.10.

Corollary 78 (SQ Lower Bound for Decision Variant of Distributional k-PC) *For any constant $\delta > 0$, suppose $k \leq n^{1/2-\delta}$. Let $D = \mathrm{Unif}(\{0,1\}^n)$ and let $\mathcal{D}$ be the set of all planted bipartite k-PC distributions (one for each clique position). Any SQ algorithm that solves the hypothesis testing problem between $\mathcal{D}$ and D with probability better than $2/3$ requires $\Omega(n^2/k^2)$ queries to the Unbiased Oracle.*

A similar statement holds for VSTAT. There is a $t = n^{\Omega(\log n)}$ such that any randomized SQ algorithm that solves the hypothesis testing problem between $\mathcal{D}$ and D with probability better than $2/3$ requires at least t queries to $VSTAT(n^{2-\delta}/k^2)$.

We conclude this section by outlining how to extend these lower bounds to distributional versions of k-BPC and BPC and why the statistical query model is not suitable to deduce hardness of problems that are implicitly tensor or hypergraph problems such as k-HPC.

Extending these SQ Lower Bounds. Extending to the bipartite case is straightforward and follows by replacing the probability of including each right vertex from k/n to k_m/m where $k_m = O(m^{1/2-\delta})$. This causes the upper bound in Lemma 74 to become $\chi_D(D_S, D_T) = |\langle \widehat{D}_S, \widehat{D}_T \rangle_D| \leq 2^{|S\cap T|} k_m^2/m^2$. Similarly, the upper bound in Lemma 75 becomes $2^{\ell+3}k_m^2/m^2$, the relevant statistical dimension becomes $SDA(\mathcal{Z}, 2^{\ell+3}k_m^2/n_m^2, \left(\frac{n}{k}\right)^{-k}) \geq n^{2\ell\delta}/8$ and the query lower bound in the final corollary becomes $\Omega(m^2/k_m^2 \log n) = \tilde{\Omega}(m^{1+2\delta})$ which yields the desired lower bound for k-BPDS. The lower bound for BPDS follows by the same extension to the ordinary PC lower bound in Feldman et al. (2013).

Hypergraph PC and SQ Lower Bounds. A key component of formulating SQ lower bounds is devising a distributional version of the problem with analogous limits in the SQ model. While there was a natural bipartite extension for PC, for hypergraph PC, such an extension does not seem to exist. Treating slices as individual samples yields a problem with statistical query algorithms that can detect a planted clique outside of polynomial time. Consider the function that given a slice, searches for a clique of size k in the induced $(s-1)$-uniform hypergraph on the neighbors of the vertex corresponding to the slice, outputting 1 if such a clique is found. Without a planted clique, the probability a slice contains such a clique is exponentially small, while it is k/n if there is a planted clique. An alternative is to consider individual entries as samples, but this discards the hypergraph structure of the problem entirely.

Appendix L. Robustness, Negative Sparse PCA and Supervised Problems

In this section, we apply reductions in Part II to deduce computational lower bounds for robust sparse mean estimation, negative sparse PCA, mixtures of SLRs and robust SLR that follow from specific instantiations of the PC_ρ conjecture. Specifically, we apply the reduction k-BPDS-TO-ISGM to deduce a lower bound for RSME, the reduction BPDS-TO-NEG-SPCA to deduce a lower bound for NEG-SPCA and the reduction k-BPDS-TO-MSLR to deduce lower bounds for MSLR, USLR and RSLR. This section is primarily devoted to summarizing the implications of these reductions and making

explicit how their input parameters need to be set to deduce our lower bounds. The implications of these lower bounds and the relation between them and algorithms was previously discussed in Section 3. In cases where the discussion in Section 3 was not exhaustive, such as the details of starting with different hardness assumptions, the number theoretic condition (T) or the adversary implied by our reductions for RSLR, we include omitted details in this section.

All lower bounds that will be shown in this section are *computational lower bounds* in the sense introduced in the beginning of Section 3. To deduce our computational lower bounds from reductions, it suffices to verify the three criteria in Condition E.1. We remark that this section is technical due to the number-theoretic constraints imposed by the prime number r in our reductions. However, these technical details are tangential to the primary focus of the paper, which is reduction techniques.

L.1. Robust Sparse Mean Estimation

We first observe that the instances of ISGM output by the reduction k-BPDS-TO-ISGM are instances of RSME in Huber's contamination model. Let r be a prime number and $\epsilon \geq 1/r$. It then follows that a sample from $\text{ISGM}_D(n, k, d, \mu, 1/r)$ is of the form

$$\text{MIX}_\epsilon\left(\mathcal{N}(\mu \cdot \mathbf{1}_S, I_d), \mathcal{D}_O\right)^{\otimes n} \quad \text{where} \quad \mathcal{D}_O = \text{MIX}_{\epsilon^{-1}r^{-1}}\left(\mathcal{N}(\mu \cdot \mathbf{1}_S, I_d), \mathcal{N}(\mu' \cdot \mathbf{1}_S, I_d)\right)$$

for some possibly random S with $|S| = k$ and where $(1 - r^{-1})\mu + r^{-1} \cdot \mu' = 0$. Note that this is a distribution in the composite hypothesis H_1 of $\text{RSME}(n, k, d, \tau, \epsilon)$ in Huber's contamination model with outlier distribution $\mathcal{D}_O$ and where $\tau = \|\mu \cdot \mathbf{1}_S\|_2 = \mu\sqrt{k}$. This observation and the discussion in Section E.2 yields that it suffices to exhibit a reduction to ISGM to show the lower bound for RSME in Theorem 4.

We now discuss the condition (T) and the number-theoretic constraint arising from applying Theorem 46 to prove Theorem 4. As mentioned in Section B.3, while this condition does not restrict our computational lower bound for RSME in the main regime of interest where $\epsilon^{-1} = n^{o(1)}$, it also can be removed using the design matrices $R_{n,\epsilon}$ in place of $K_{r,t}$. Despite this, we introduce the condition (T) in this section as it will be a necessary condition in subsequent lower bounds in Part III.

As discussed in Section I, the prime power r^t in k-BPDS-TO-ISGM is intended to be a fairly close approximation to each of $k_n, \sqrt{n}$ and $\sqrt{N}$. We will now see that in order to show tight computational lower bounds for RSME, this approximation needs to be very close to asymptotically exact, leading to the technical condition (T). First note that the level of signal μ produced by the reduction k-BPDS-TO-ISGM is

$$\mu \leq \frac{\delta}{2\sqrt{6\log(k_n m r^t) + 2\log(p-q)^{-1}}} \cdot \frac{1}{\sqrt{r^t(r-1)(1+(r-1)^{-1})}} = \tilde{\Theta}\left(r^{-(t+1)/2}\right)$$

where $\delta = \Theta(1)$ and the estimate above holds whenever p and q are constants. Therefore the corresponding τ is given by $\tau = \mu\sqrt{k} = \tilde{O}(k^{1/2}r^{-(t+1)/2})$. Furthermore, in Theorem 46, the output number of samples N is constrained to satisfy that $N = o(k_n r^t)$ and $n = O(k_n r^t)$. Combining this with the fact that in order to be starting with a hard k-BPDS instance, we need $k_n = o(\sqrt{n})$ to hold, it is straightforward to see that these constraints together require that $N = o(r^{2t})$. If this is close to tight with $N = \tilde{\Theta}(r^{2t})$, the computational lower bound condition on τ becomes

$$\tau = \tilde{O}\left(k^{1/2}r^{-(t+1)/2}\right) = \tilde{\Theta}\left(k^{1/2}\epsilon^{1/2}N^{-1/4}\right)$$

where we also use the fact that $\epsilon = \Theta(1/r)$. Note that this corresponds exactly to the desired computational lower bound of $N = \tilde{o}(k^2\epsilon^2/\tau^4)$. Furthermore, if instead $N = \tilde{\Theta}(a^{-1}r^{2t})$ for some $a = \omega(1)$, then the lower bound we show degrades to $N = \tilde{o}(k^2\epsilon^2/a\tau^4)$, and is suboptimal by a factor $a = \omega(1)$. Thus ideally we would like the pair of parameters (N, r) to be such that there infinitely many N with something like $N = \tilde{\Theta}(r^{2t})$ true for some positive integer $t \in \mathbb{N}$. This leads exactly to the condition (T) below.

Definition 79 (Condition (T)) *Suppose that (N, r) is a pair of parameters with $N \in \mathbb{N}$ and $r = r(N)$ is non-decreasing. The pair (N, r) satisfies* (T) *if either $r = N^{o(1)}$ as $N \to \infty$ or if $r = \tilde{\Theta}(N^{1/t})$ where $t \in \mathbb{N}$ is a constant even integer.*

The key property arising from condition (T) is captured in the following lemma.

Lemma 80 (Property of (T)) *Suppose that (N, r) satisfies* (T) *and let $r' = r'(N)$ be any non-decreasing positive integer parameter satisfying that $r' = \tilde{\Theta}(r)$. Then there are infinitely many values of N with the following property: there exists $s \in \mathbb{N}$ such that $\sqrt{N} = \tilde{\Theta}\left((r')^s\right)$.*

Proof If $r = \tilde{\Theta}(N^{1/t})$ where $t \in \mathbb{N}$ is a constant even integer, then this property is satisfied trivially by taking $s = t/2$. Now suppose that $r = N^{o(1)}$ and note that this also implies that $r' = N^{o(1)}$. Now consider the function

$$f(N) = \frac{\log N}{2 \log r'(N)}$$

Since $r' = N^{o(1)}$, it follows that $f(N) \to \infty$ as $N \to \infty$. Suppose that N is sufficiently large so that $f(N) > 1$. Note that, for each N, either $r'(N+1) \geq r'(N) + 1$ or $r'(N+1) = r'(N)$. If $r'(N+1) = r'(N)$, then $f(N+1) > f(N)$. If $r'(N+1) \geq r'(N) + 1$, then

$$\frac{f(N+1)}{f(N)} \leq \frac{g(N)}{g(r'(N))} \quad \text{where} \quad g(x) = \frac{\log(x+1)}{\log x}$$

Note that $g(x)$ is a decreasing function of x for $x \geq 2$. Since $f(N) > 1$, it follows that $r'(N) < N$ and hence the above inequality implies that $f(N+1) < f(N)$. Summarizing these observations, every time $f(N)$ increases it must follow that $r'(N+1) = r'(N)$. Fix a sufficiently large positive integer s and consider the first N for which $f(N) \geq s$. It follows by our observation that $r'(N) = r'(N-1)$ and furthermore that $f(N-1) < s$. This implies that $N - 1 < r'(N)^{2s}$ and $N \geq r'(N)^{2s}$. Since $r'(N)$ is a positive integer, it then must follow that $N = r'(N)^{2s}$. Since such an N exists for every sufficiently large s, this completes the proof of the lemma. ■

This condition (T) will arise in a number of others problems that we map to, including robust SLR and dense stochastic block models, for a nearly identical reason. We now formally prove Theorem 4. All remaining proofs in this section will be of a similar flavor and where details are similar, we only sketch them to avoid redundancy.

Theorem 4 (Lower Bounds for RSME) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $\epsilon < 1/2$ is such that (n, ϵ^{-1}) satisfies* (T)*, then the k-*BPC *conjecture or k-*BPDS *conjecture for constant $0 < q < p \leq 1$ both imply that there is a computational lower bound for* RSME$(n, k, d, \tau, \epsilon)$ *at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

Proof To prove this theorem, we will to show that Theorem 46 implies that k-BPDS-TO-ISGM fills out all of the possible growth rates specified by the computational lower bound $n = \tilde{o}(k^2\epsilon^2/\tau^4)$ and the other conditions in the theorem statement. As discussed earlier in this section, it suffices to reduce in total variation to ISGM$(n, k, d, \mu, 1/r)$ where $1/r \leq \epsilon$ and $\mu = \tau/\sqrt{k}$.

Fix a constant pair of probabilities $0 < q < p \leq 1$ and any sequence of parameters $(n, k, d, \tau, \epsilon)$ all of which are implicitly functions of n such that (n, ϵ^{-1}) satisfies (T) and $(n, k, d, \tau, \epsilon)$ satisfy the conditions

$$n \leq c \cdot \frac{k^2\epsilon^2}{\tau^4 \cdot (\log n)^{2+2c'}} \quad \text{and} \quad wk^2 \leq d$$

for sufficiently large n, an arbitrarily slow-growing function $w = w(n) \to \infty$ at least satisfying that $w(n) = n^{o(1)}$, a sufficiently small constant $c > 0$ and a sufficiently large constant $c' > 0$. In order to fulfill the criteria in Condition E.1, we now will specify:

1. a sequence of parameters (M, N, k_M, k_N, p, q) such that the k-BPDS instance with these parameters is hard according to Conjecture 3; and

2. a sequence of parameters $(n', k, d, \tau, \epsilon)$ with a subsequence that satisfies three conditions: (2.1) the parameters on the subsequence are in the regime of the desired computational lower bound for RSME; (2.2) they have the same growth rate as $(n, k, d, \tau, \epsilon)$ on this subsequence; and (2.3) such that RSME with the parameters on this subsequence can be produced by the reduction k-BPDS-TO-ISGM with input k-BPDS(M, N, k_M, k_N, p, q).

By the discussion in Section E.2, this would be sufficient to show the desired computational lower bound. We choose these parameters as follows:

- let r be a prime with $r \geq \epsilon^{-1}$ and $r \leq 2\epsilon^{-1}$, which exists by Bertrand's postulate and can be found in poly$(\epsilon^{-1}) \leq$ poly(n) time;
- let t be such that r^t is the closest power of r to $\sqrt{n}$, let $n' = \lfloor w^{-2}r^{2t} \rfloor$, let $k_N = \lfloor \sqrt{n'} \rfloor$ and let $N = wk_N^2 \leq k_N r^t$; and
- set $\mu = \tau/\sqrt{k}$, $k_M = k$ and $M = wk^2$.

The given inequality and parameter settings above rearrange to the following condition on n':

$$n' \leq w^{-2}r^{2t} = O\left(\frac{r^{2t}}{n} \cdot \frac{k^2\epsilon^2}{\tau^4 \cdot (\log n)^{2+2c'}}\right)$$

Furthermore, the given inequality yields the constraint on μ that

$$\mu = \tau \cdot k^{-1/2} \leq \frac{c^{1/4}\epsilon^{1/2}}{n^{1/4}(\log n)^{(1+c')/2}} = \Theta\left(\frac{r^{t/2}}{n^{1/4}} \cdot \frac{1}{\sqrt{r^{t+1}(\log n)^{1+c'}}}\right)$$

As long as $\sqrt{n} = \tilde{\Theta}(r^t)$ then: (2.1) the inequality above on n' would imply that $(n', k, d, \tau, \epsilon)$ is in the desired hard regime; (2.2) n and n' have the same growth rate since $w = n^{o(1)}$; and (2.3) taking c' large enough would imply that μ satisfies the conditions needed to apply Theorem 46 to yield the desired reduction. By Lemma 80, there is an infinite subsequence of the input parameters such that $\sqrt{n} = \tilde{\Theta}(r^t)$. This verifies the three criteria in Condition E.1. Following the argument in Section E.2, Lemma 14 now implies the theorem. ■

As alluded to in Section B.3, replacing $K_{r,t}$ with $R_{n,\epsilon}$ in the applications of dense Bernoulli rotations in k-BPDS-TO-ISGM removes condition (T) from this lower bound. Specifically, applying k-BPDS-TO-ISGM$_R$ and Corollary 49 in place of k-BPDS-TO-ISGM and replacing the dimension r^t with L in the argument above yields the lower bound shown below. Note that condition (T) in Theorem 4 is replaced by the looser requirement that $\epsilon = \tilde{\Omega}(n^{-1/2})$. As discussed at the end of Section I.1, this requirement arises from the condition $\epsilon \gg L^{-1} \log L$ in Corollary 49. We remark that the condition $\epsilon = \tilde{\Omega}(n^{-1/2})$ is implicit in (T) and hence the following corollary is strictly stronger than Theorem 4.

Corollary 81 (Lower Bounds for RSME without Condition (T)) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $\epsilon < 1/2$ is such that $\epsilon = \tilde{\Omega}(n^{-1/2})$, then the k-BPC conjecture or k-BPDS conjecture for constant $0 < q < p \le 1$ both imply that there is a computational lower bound for RSME$(n, k, d, \tau, \epsilon)$ at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

We remark that only assuming the k-PC conjecture also yields hardness for RSME. In particular k-PC can be mapped to the asymmetric bipartite case by considering the bipartite subgraph with $k/2$ parts on one size and $k/2$ on the other. Showing hardness for RSME from k-PC then reduces to the hardness yielded by k-BPC with $M = N$. Examining this restricted setting in the theorem above and passing through an analogous argument yields a computational lower bound at the slightly suboptimal rate

$$n = \tilde{o}\left(k^2\epsilon/\tau^2\right) \quad \text{as long as} \quad \tau^2 \log n = o(\epsilon)$$

When $(\log n)^{-O(1)} \lesssim \epsilon \lesssim 1/\log n$, then the optimal k-to-k^2 gap is recovered up to polylog(n) factors by this result.

L.2. Negative Sparse PCA

In this section, we deduce Theorem 7 on the hardness of NEG-SPCA using the reduction BPDS-TO-NEG-SPCA and Theorem 43. Because this reduction does not bear the number-theoretic considerations of the reduction to RSME, this proof will be substantially more straightforward.

Theorem 7 (Lower Bounds for NEG-SPCA) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the BPC or BPDS conjecture for constant $0 < q < p \le 1$ both imply conjecture implies a computational lower bound for NEG-SPCA(n, k, d, θ) at all levels of signal $\theta = \tilde{o}(\sqrt{k^2/n})$.*

Proof We show that Theorem 43 implies that BPDS-TO-NEG-SPCA fills out all of the possible growth rates specified by the computational lower bound $\theta = \tilde{o}(\sqrt{k^2/n})$ and the other conditions in the theorem statement. Fix a constant pair of probabilities $0 < q < p \le 1$ and a sequence of parameters (n, k, d, θ) all of which are implicitly functions of n such that

$$\theta \le cw^{-1} \cdot \sqrt{\frac{k^2}{n(\log n)^2}}, \quad wk \le n^{1/6} \quad \text{and} \quad wk^2 \le d$$

for sufficiently large n, an arbitrarily slow-growing function $w = w(n) \to \infty$ where $w(n) = n^{o(1)}$ and a sufficiently small constant $c > 0$. In order to fulfill the criteria in Condition E.1, we now will

specify: a sequence of parameters (M, N, k_M, k_N, p, q) such that the BPDS instance with these parameters is hard according to Conjecture 3, and such that NEG-SPCA with the parameters (n, k, d, θ) can be produced by the reduction BPDS-TO-NEG-SPCA applied to BPDS(M, N, k_M, k_N, p, q). These parameters along with the internal parameter τ of the reduction can be chosen as follows:

- let $N = n$, $k_N = w^{-1}\sqrt{n}$, $k_M = k$ and $M = wk^2$; and
- let $\tau > 0$ be such that
$$\tau^2 = \frac{4n\theta}{k_N k(1-\theta)}$$

It is straightforward to verify that the inequality above upper bounding θ implies that $\tau \le 4c/\sqrt{\log n}$ and thus satisfies the condition on τ needed to apply Lemma 39 and Theorem 43 for a sufficiently small $c > 0$. Furthermore, this setting of τ yields

$$\theta = \frac{\tau^2 k_N k}{4n + \tau^2 k_N k}$$

Furthermore, note that $d \ge M$ and $n \gg M^3$ by construction. Applying Theorem 43 now verifies the desired property above. This verifies the criteria in Condition E.1 and, following the argument in Section E.2, Lemma 14 now implies the theorem. ■

We remark the the constraint $k = o(n^{1/6})$, as mentioned in Section B.7, is a technical condition that we believe should not be necessary for the theorem to hold. This is similar to the constraint arising in the strong reduction to sparse PCA given by CLIQUE-TO-WISHART in Brennan and Bresler (2019). In CLIQUE-TO-WISHART, the random matrix comparison between Wishart and GOE produced the technical condition that $k = o(n^{1/6})$ in a similar manner to how our comparison result between Wishart and inverse Wishart produces the same constraint here. We also remark that the reduction CLIQUE-TO-WISHART can be used here to yield the same hardness for NEG-SPCA as in Theorem 7 based only on the PC conjecture. This is achieved by the reduction that maps from PC to sparse PCA with $d = wk^2$ as a first step using CLIQUE-TO-WISHART and then uses the second step of BPDS-TO-NEG-SPCA to map to NEG-SPCA.

L.3. Mixtures of Sparse Linear Regressions and Robustness

In this section, we deduce Theorems 82, 8 and 9 on the hardness of unsigned, mixtures of and robust sparse linear regression, all using the reduction k-BPDS-TO-MSLR with different parameters (r, ϵ) and Theorem 52. We begin by showing bounds for USLR(n, k, d, τ).

We first make the following simple but important observation. Note that a single sample from USLR is of the form $y = |\tau \cdot \langle v_S, X\rangle + \mathcal{N}(0,1)|$, which has the same distribution as $|y'|$ where $y' = \tau r \cdot \langle v_S, X\rangle + \mathcal{N}(0,1)$ and r is an independent Rademacher random variable. Note that y' is a sample from MSLR$_D(n, k, d, \gamma, 1/2)$ with $\gamma = \tau$. Thus to show a computational lower bound for USLR(n, k, d, τ), it suffices to show a lower bound for MSLR(n, k, d, τ).

Theorem 82 (Lower Bounds for USLR) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the k-BPC or k-BPDS conjecture for constant $0 < q < p \le 1$ both imply that there is a computational lower bound for* USLR(n, k, d, τ) *at all sample complexities $n = \tilde{o}(k^2/\tau^4)$.*

Proof To prove this theorem, we will show that Theorem 52 implies that k-BPDS-TO-MSLR applied with $r = 2$ fills out all of the possible growth rates specified by the computational lower bound $n = \tilde{o}(k^2/\tau^4)$ and the other conditions in the theorem statement. As mentioned above, it suffices to reduce in total variation to $\text{MSLR}(n, k, d, \tau)$. Fix a constant pair of probabilities $0 < q < p \leq 1$ and any sequence of parameters (n, k, d, τ) all of which are implicitly functions of n with

$$n \leq c \cdot \frac{k^2}{w^2 \cdot \tau^4 \cdot (\log n)^4}, \quad wk \leq n^{1/6} \quad \text{and} \quad wk^2 \leq d$$

for sufficiently large n, an arbitrarily slow-growing function $w = w(n) \to \infty$ and a sufficiently small constant $c > 0$. In order to fulfill the criteria in Condition E.1, we now will specify: a sequence of parameters (M, N, k_M, k_N, p, q) such that the k-BPDS instance with these parameters is hard according to Conjecture 3, and such that MSLR with the parameters $(n, k, d, \tau, 1/2)$ can be produced by the reduction k-BPDS-TO-MSLR applied with $r = 2$ to $\text{BPDS}(M, N, k_M, k_N, p, q)$. By the discussion in Section E.2, this would be sufficient to show the desired computational lower bound. We choose these parameters as follows:

- let t be such that 2^t is the smallest power of two greater than $w\sqrt{n}$, let $k_N = \lfloor \sqrt{n} \rfloor$ and let $N = wk_N^2 \leq k_N 2^t$; and
- set $k_M = k$ and $M = wk^2$.

Now note that τ^2 is upper bounded by

$$\tau^2 \leq \frac{c^{1/2} \cdot k}{wn^{1/2} \cdot (\log n)^2} = O\left(\frac{k_N k_M}{N \log(MN)}\right)$$

Furthermore, we have that

$$\tau^2 \leq \frac{c^{1/2} \cdot k}{wn^{1/2} \cdot (\log n)^2} = \Theta\left(\frac{k_M}{2^{t+1} \log(k_N M \cdot 2^t) \log n}\right)$$

Therefore τ satisfies the conditions needed to apply Theorem 52 for a sufficiently small $c > 0$. Also note that $n \gg M^3$ and $d \geq M$ by construction. Applying Theorem 52 now verifies the desired property above. This verifies the criteria in Condition E.1 and, following the argument in Section E.2, Lemma 14 now implies the theorem. ■

The proof of the theorem above also directly implies Theorem 8. This yields our main computational lower bounds for MSLR, which are stated below.

Theorem 8 (Lower Bounds for MSLR) *If k, d and n are polynomial in each other, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the k-BPC or k-BPDS conjecture for constant $0 < q < p \leq 1$ both imply that there is a computational lower bound for $\text{MSLR}(n, k, d, \tau)$ at all sample complexities $n = \tilde{o}(k^2/\tau^4)$.*

Now observe that the instances of MSLR output by the reduction k-BPDS-TO-MSLR applied with $r > 2$ are instances of RSLR in Huber's contamination model. Let r be a prime number and $\epsilon \geq 1/r$. Also let $X \sim \mathcal{N}(0, I_d)$ and $y = \tau \cdot \langle v_S, X \rangle + \eta$ where $\eta \sim \mathcal{N}(0, 1)$ where $|S| = k$. By Definition 51, $\text{MSLR}_D(n, k, d, \tau, 1/r)$ is of the form

$$\text{MIX}_\epsilon \left(\mathcal{L}(X, y), \mathcal{D}_O\right)^{\otimes n} \quad \text{where} \quad \mathcal{D}_O = \text{MIX}_{\epsilon^{-1}r^{-1}} \left(\mathcal{L}(X, y), \mathcal{L}'\right)$$

for some possibly random S with $|S| = k$ and where $\mathcal{L}'$ denotes the distribution on pairs (X, y) that are jointly Gaussian with mean zero and $(d+1) \times (d+1)$ covariance matrix

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{Xy} \\ \Sigma_{yX} & \Sigma_{yy} \end{bmatrix} = \begin{bmatrix} I_d + \frac{(a^2-1)\gamma^2}{1+\gamma^2} \cdot v_S v_S^\top & -a\gamma \cdot v_S \\ -a\gamma \cdot v_S^\top & 1+\gamma^2 \end{bmatrix}$$

This yields a very particular construction of an adversary in Huber's contamination model, which we show in the next theorem yields a computational lower bound for RSLR. With the observations above, the proof of this theorem is similar to that of Theorem 4 and is deferred to Appendix R.2.

Theorem 83 (Lower Bounds for RSLR with Condition (T)) *If k, d and n are polynomial in each other, $\epsilon < 1/2$ is such that (n, ϵ^{-1}) satisfies* (T)*, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the k-BPC conjecture or k-BPDS conjecture for constant $0 < q < p \leq 1$ both imply that there is a computational lower bound for* RSLR$(n, k, d, \tau, \epsilon)$ *at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

Our main computational lower bound for RSLR follows from the same argument applied to the reduction k-BPDS-TO-MSLR$_R$ instead of k-BPDS-TO-MSLR and using Corollary 57 instead of Theorem 52. As in Corollary 81, this replaces condition (T) with the weaker condition that $\epsilon = \tilde{\Omega}(n^{-1/2})$.

Theorem 9 (Lower Bounds for RSLR) *If k, d and n are polynomial in each other, $\epsilon < 1/2$ is such that $\epsilon = \tilde{\Omega}(n^{-1/2})$, $k = o(\sqrt{d})$ and $k = o(n^{1/6})$, then the k-BPC conjecture or k-BPDS conjecture for constant $0 < q < p \leq 1$ both imply that there is a computational lower bound for* RSLR$(n, k, d, \tau, \epsilon)$ *at all sample complexities $n = \tilde{o}(k^2\epsilon^2/\tau^4)$.*

Appendix M. Community Recovery and Partition Models

In this section, we devise several reductions based on BERN-ROTATIONS and TENSOR-BERN-ROTATIONS using the design matrices and tensors from Section G to reduce from k-PC, k-PDS, k-BPC and k-BPDS to dense stochastic block models, hidden partition models and semirandom planted dense subgraph. These reductions are briefly outlined in Section C.3.

Furthermore, the heuristic presented at the end of Section C.3 predicts the computational barriers for the problems in this section. The ℓ_2 norm of the matrix $\mathbb{E}[X]$ corresponding to a k-PC instance is $\Theta(k)$, which is just below $\tilde{\Theta}(\sqrt{n})$ when this k-PC is near its computational barrier. Furthermore, it can be verified that the ℓ_2 norm of the matrices $\mathbb{E}[X]$ corresponding to the problems in this section are:

- If $\gamma = P_{11} - P_0$ in the ISBM notation of Section B.4, then a direct calculation yields that the ℓ_2 norm corresponding to ISBM is $\Theta(n\gamma/k)$.
- In GHPM and BHPM, the corresponding ℓ_2 norm can be verified to be $\Theta(K\gamma\sqrt{r})$.
- In our adversarial construction for SEMI-CR, the corresponding ℓ_2 norm is $\Theta(k\gamma)$ where $\gamma = P_1 - P_0$.

Following the heuristic, setting these equal to $\tilde{\Theta}(\sqrt{n})$ yields the predicted computational barriers of $\gamma^2 = \tilde{\Theta}(k^2/n)$ in ISBM, $\gamma^2 = \tilde{\Theta}(n/rK^2)$ in GHPM and BHPM and $\gamma^2 = \tilde{\Theta}(n/k^2)$ in SEMI-CR. We now present our reduction to ISBM.

M.1. Dense Stochastic Block Models with Two Communities

We begin by recalling the definition of the imbalanced 2-block stochastic block model from Section B.4.

Definition 84 (Imbalanced 2-Block Stochastic Block Model) *Let k and n be positive integers such that k divides n. The distribution* $\text{ISBM}_D(n, k, P_{11}, P_{12}, P_{22})$ *over n-vertex graphs G is sampled by first choosing an (n/k)-subset $C \subseteq [n]$ uniformly at random and sampling the edges of G independently with the following probabilities*

$$\mathbb{P}\left[\{i,j\} \in E(G)\right] = \begin{cases} P_{11} & \text{if } i,j \in C \\ P_{12} & \text{if exactly one of } i,j \text{ is in } C \\ P_{22} & \text{if } i,j \in [n] \backslash C \end{cases}$$

Given a subset $C \subseteq [n]$ of size n/k, we let $\text{ISBM}_D(n, C, P_{11}, P_{12}, P_{22})$ denote ISBM as defined above conditioned on the latent subset C. As discussed in Section E.3, this naturally leads to a composite hypothesis testing problem between

$$H_0 : G \sim \mathcal{G}(n, P_0) \quad \text{and} \quad H_1 : G \sim \text{ISBM}_D(n, k, P_{11}, P_{12}, P_{22})$$

where P_0 is any edge density in $(0, 1)$. This section is devoted to showing reductions from k-PDS and k-PC to ISBM formulated as this hypothesis testing problem. In particular, we will focus on P_{11}, P_{12}, P_{22} and P_0 all of which are bounded away from 0 and 1 by a constant, and which satisfy that

$$P_0 = \frac{1}{k} \cdot P_{11} + \left(1 - \frac{1}{k}\right) P_{12} = \frac{1}{k} \cdot P_{12} + \left(1 - \frac{1}{k}\right) P_{22} \tag{11}$$

These two constraints allow P_{11}, P_{12}, P_{22} to be reparameterized in terms of a signal parameter γ as

$$P_{11} = P_0 + \gamma, \quad P_{12} = P_0 - \frac{\gamma}{k-1} \quad \text{and} \quad P_{22} = P_0 + \frac{\gamma}{(k-1)^2} \tag{12}$$

There are two main reasons why we restrict to the parameter regime enforced by the density constraints in (11) – it creates a model with nearly uniform expected degrees and which is a mean-field analogue of recovering the first community in the k-block stochastic block model.

- *Nearly Uniform Expected Degrees*: Observe that, conditioned on C, the expected degree of a vertex $i \in [n]$ in $\text{ISBM}(n, k, P_{11}, P_{12}, P_{22})$ is given by

$$\mathbb{E}\left[\deg(i) | C\right] = \begin{cases} \left(\frac{n}{k} - 1\right) \cdot P_{11} + \frac{n(k-1)}{k} \cdot P_{12} & \text{if } i \in C \\ \frac{n}{k} \cdot P_{12} + \left(\frac{n(k-1)}{k} - 1\right) \cdot P_{22} & \text{if } i \in [n] \backslash C \end{cases}$$

  Thus the density constraints in (11) ensure that these differ by at most 1 from each other and from $(n-1)P_0$. Thus all of the vertices in $\text{ISBM}(n, k, P_{11}, P_{12}, P_{22})$ and the H_0 model $\mathcal{G}(n, P_0)$ have approximately the same expected degree. This precludes simple degree and total edge thresholding tests that are optimal in models of single community detection that are not degree-corrected. As discussed in Section B.6, the planted dense subgraph model has a detection threshold that differs from the conjectured Kesten-Stigum threshold for recovery of the planted dense subgraph. Thus to obtain computational lower bounds for a hypothesis testing problem that give tight recovery lower bounds, calibrating degrees is crucial. The main result of this section can be viewed as showing approximate degree correction is sufficient to obtain the Kesten-Stigum threshold for ISBM through a reduction from k-PDS and k-PC.

- *Mean-Field Analogue of First Community Recovery in k-SBM:* As discussed in Section B.4, the imbalanced 2-block stochastic block model $\text{ISBM}_D(n, k, P_{11}, P_{12}, P_{22})$ is roughly a mean-field analogue of recovering the first community C_1 in a k-block stochastic block model. More precisely, consider a graph G wherein the vertex set $[n]$ is partitioned into k latent communities $C_1, C_2, \ldots, C_k$ each of size n/k and edges are then included in the graph G independently such that intra-community edges appear with probability p while inter-community edges appear with probability $q < p$. The distribution $\text{ISBM}_D(n, k, P_{11}, P_{12}, P_{22})$ can be viewed as a mean-field analogue of recovering a first community $C = C_1$ in the k-block model, when

$$P_{11} = p, \quad P_{12} = q \quad \text{and} \quad P_{22} = \frac{1}{k-1} \cdot p + \left(1 - \frac{1}{k-1}\right) q$$

  Here, P_{22} approximately corresponds to the average edge density on the subgraph of the k-block model restricted to $[n] \backslash C_1$. This analogy between ISBM and k-SBM is also why we choose to parameterize ISBM in terms of k rather than the size n/k of C.

As discussed in Section B.4, if $k = o(\sqrt{n})$, the conjectured recovery threshold for efficient recovery in k-SBM is the Kesten-Stigum threshold of

$$\frac{(p-q)^2}{q(1-q)} \gtrsim \frac{k^2}{n}$$

while the statistically optimal rate of recovery is when this level of signal is instead $\tilde{\Omega}(k^4/n^2)$. Furthermore, the information-theoretic threshold and conjectured computational barrier are the same for ISBM in the regime defined by (11). Parameterizing ISBM in terms of γ as in (12), the Kesten-Stigum threshold can be expressed as $\gamma^2 = \tilde{\Omega}(k^2/n)$. The objective of this section is give a reduction from k-PDS to ISBM in the dense regime with $\min\{P_0, 1-P_0\} = \Omega(1)$ up to the Kesten-Stigum threshold.

The first reduction of this section k-PDS-TO-ISBM is shown in Figure 12 and maps to the case where $P_0 = 1/2$ and (12) is only approximately true. In a subsequent corollary, a simple modification of this reduction will map to all P_0 with $\min\{P_0, 1 - P_0\} = \Omega(1)$ and show (12) holds exactly. The following theorem establishes the approximate Markov transition properties of k-PDS-TO-ISBM. The proof of this theorem follows a similar structure to the proof of Theorem 46. Recall that $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$ denotes the standard normal CDF.

Theorem 85 (Reduction to ISBM) *Let N be a parameter and $r = r(N) \geq 2$ be a prime number. Fix initial and target parameters as follows:*

- Initial k-BPDS Parameters: *vertex count N, subgraph size $k = o(N)$ dividing N, edge probabilities $0 < q < p \leq 1$ with $\min\{q, 1-q\} = \Omega(1)$ and $p - q \geq N^{-O(1)}$, and a partition E of $[N]$.*
- Target ISBM Parameters: (n, r) *where $\ell = \frac{r^t - 1}{r-1}$ and $n = kr\ell$ for some parameter $t = t(N) \in \mathbb{N}$ satisfying that that*

$$m \leq kr^t \leq kr\ell \leq \text{poly}(N)$$

  where m is the smallest multiple of k larger than $\left(\frac{p}{Q} + 1\right) N$ and where

$$Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}} (\sqrt{q} - 1)$$

- Target ISBM Edge Strengths: (P_{11}, P_{12}, P_{22}) *given by*

$$P_{11} = \Phi\left(\frac{\mu(r-1)^2}{r^{t+1}}\right), \quad P_{12} = \Phi\left(-\frac{\mu(r-1)}{r^{t+1}}\right) \quad \text{and} \quad P_{22} = \Phi\left(\frac{\mu}{r^{t+1}}\right)$$

where $\mu \in (0,1)$ *satisfies that*

$$\mu \leq \frac{1}{2\sqrt{6\log(kr\ell) + 2\log(p-Q)^{-1}}} \cdot \min\left\{\log\left(\frac{p}{Q}\right), \log\left(\frac{1-Q}{1-p}\right)\right\}$$

Let $\mathcal{A}(G)$ *denote* k-PDS-TO-ISBM *applied to the graph* G *with these parameters. Then* $\mathcal{A}$ *runs in* $\mathrm{poly}(N)$ *time and it follows that*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}_E(N,k,p,q)\right), \text{ISBM}_D(n,r,P_{11},P_{12},P_{22})\right) = O\left(\frac{k}{\sqrt{N}} + e^{-\Omega(N^2/km)} + (kr\ell)^{-1}\right)$$
$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}(N,q)\right), \mathcal{G}(n,1/2)\right) = O\left(e^{-\Omega(N^2/km)} + (kr\ell)^{-1}\right)$$

To prove this theorem, we begin by proving a lemma analyzing the dense Bernoulli rotations step of k-PDS-TO-ISBM. Define $v_{S,F',F''}(M)$ as in Section I.1. The proof of the next lemma follows similar steps to the proof of Lemma 47.

Lemma 86 (Bernoulli Rotations for ISBM) *Let* F' *and* F'' *be a fixed partitions of* $[kr^t]$ *and* $[kr\ell]$ *into* k *parts of size* r^t *and* $r\ell$*, respectively, and let* $S \subseteq [kr^t]$ *where* $|S \cap F_i'| = 1$ *for each* $1 \leq i \leq k$*. Let* $\mathcal{A}_3$ *denote Step 3 of* k-PDS-TO-ISBM *with input* M_{PD2} *and output* M_{R}*. Suppose that* p, Q *and* μ *are as in Theorem 85, then it follows that*

$$d_{TV}\Big(\mathcal{A}_3\left(\mathcal{M}_{[kr^t]\times[kr^t]}\left(S\times S, \text{Bern}(p), \text{Bern}(Q)\right)\right), \mathcal{L}\left(\frac{\mu(r-1)}{r}\cdot v_{S,F',F''}(K_{r,t})v_{S,F',F''}(K_{r,t})^\top + \mathcal{N}(0,1)^{\otimes kr\ell\times kr\ell}\right)\Big) = O\left((kr\ell)^{-1}\right)$$
$$d_{TV}\left(\mathcal{A}_3\left(\text{Bern}(Q)^{\otimes kr^t\times kr^t}\right), \mathcal{N}(0,1)^{\otimes kr\ell\times kr\ell}\right) = O\left((kr\ell)^{-1}\right)$$

Proof First consider the case where $M_{\text{PD2}} \sim \mathcal{M}_{[kr^t]\times[kr^t]}\left(S\times S, \text{Bern}(p), \text{Bern}(Q)\right)$. Observe that the submatrices of M_{PD2} are distributed as follows

$$(M_{\text{PD2}})_{F_i',F_j'} \sim \text{PB}\left(F_i'\times F_j', (S\cap F_i', S\cap F_j'), p, Q\right)$$

and are independent. Combining upper bound on the singular values of $K_{r,t}$ in Lemma 30 with Corollary 27 implies that

$$d_{\text{TV}}\left((M_{\text{R}})_{F_i'',F_j''}, \mathcal{L}\left(\frac{\mu(r-1)}{r}\cdot(K_{r,t})_{\cdot,S\cap F_i'}(K_{r,t})_{\cdot,S\cap F_j'}^\top + \mathcal{N}(0,1)^{\otimes r\ell\times r\ell}\right)\right) = O\left(r^{2t}\cdot(kr\ell)^{-3}\right)$$

Since the submatrices $(M_{\text{R}})_{F_i'',F_j''}$ are independent, the tensorization property of total variation in Fact 15 implies that $d_{\text{TV}}\left(M_{\text{R}}, \mathcal{L}(Z)\right) = O\left(k^2r^{2t}\cdot(kr\ell)^{-3}\right) = O\left((kr\ell)^{-1}\right)$ where the submatrices $Z_{F_i'',F_j''}$ are independent and satisfy

$$Z_{F_i'',F_j''} \sim \mathcal{L}\left(\frac{\mu(r-1)}{r}\cdot(K_{r,t})_{\cdot,S\cap F_i'}(K_{r,t})_{\cdot,S\cap F_j'}^\top + \mathcal{N}(0,1)^{\otimes r\ell\times r\ell}\right)$$

Algorithm k-PDS-TO-ISBM

Inputs: k-PDS instance $G \in \mathcal{G}_N$ with dense subgraph size k that divides N, and the following parameters

- partition E of $[N]$ into k parts of size N/k, edge probabilities $0 < q < p \leq 1$
- let m be the smallest multiple of k larger than $\left(\frac{p}{Q} + 1\right) N$ where $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}} \left(\sqrt{q} - 1\right)$
- output number of vertices $n = kr\ell$ where r is a prime number r, $\ell = \frac{r^t - 1}{r-1}$ for some $t \in \mathbb{N}$ and $m \leq kr^t \leq kr\ell \leq \text{poly}(N)$
- mean parameter $\mu \in (0, 1)$ satisfying that

$$\mu \leq \frac{1}{2\sqrt{6 \log n + 2 \log(p - Q)^{-1}}} \cdot \min \left\{ \log \left(\frac{p}{Q} \right), \log \left(\frac{1-Q}{1-p} \right) \right\}$$

1. *Symmetrize and Plant Diagonals*: Compute $M_{\text{PD1}} \in \{0,1\}^{m \times m}$ with partition F of $[m]$ as $M_{\text{PD1}} \leftarrow$ TO-k-PARTITE-SUBMATRIX(G) applied with initial dimension N, partition E, edge probabilities p and q and target dimension m.
2. *Pad*: Form $M_{\text{PD2}} \in \{0,1\}^{kr^t \times kr^t}$ by embedding M_{PD1} as the upper left principal submatrix of M_{PD2} and then adding $kr^t - m$ new indices for columns and rows, with all missing entries sampled i.i.d. from Bern(Q). Let F_i' be F_i with $r^t - m/k$ of the new indices. Sample k random permutations σ_i of F_i' independently for each $1 \leq i \leq k$ and permute the indices of the rows and columns of M_{PD2} within each part F_i' according to σ_i.
3. *Bernoulli Rotations*: Let F'' be a partition of $[kr\ell]$ into k equally sized parts. Now compute the matrix $M_{\text{R}} \in \mathbb{R}^{kr\ell \times kr\ell}$ as follows:
   (1) For each $i, j \in [k]$, apply TENSOR-BERN-ROTATIONS to the matrix $(M_{\text{PD2}})_{F_i', F_j'}$ with matrix parameter $A_1 = A_2 = K_{r,t}$, rejection kernel parameter $R_{\text{RK}} = kr\ell$, Bernoulli probabilities $0 < Q < p \leq 1$, output dimension $r\ell$, $\lambda_1 = \lambda_2 = \sqrt{1 + (r-1)^{-1}}$ and mean parameter μ.
   (2) Set the entries of $(M_{\text{R}})_{F_i'', F_j''}$ to be the entries in order of the matrix output in (1).
4. *Threshold and Output*: Now construct the graph G' with vertex set $[kr\ell]$ such that for each $i > j$ with $i, j \in [kr\ell]$, we have $\{i, j\} \in E(G')$ if and only if $(M_{\text{R}})_{ij} \geq 0$. Output G' with randomly permuted vertex labels.

Figure 12: Reduction from k-partite planted dense subgraph to the dense imbalanced 2-block stochastic block model.

Note that the entries of Z are independent Gaussians each with variance 1 and Z has mean given by $\mu(1+r^{-1}) \cdot v_{S,F',F''}(K_{r,t}) v_{S,F',F''}(K_{r,t})^\top$, by the definition of $v_{S,F',F''}(K_{r,t})$. This proves the first total variation upper bound in the statement of the lemma. Now suppose that $M_{\text{PD2}} \sim \text{Bern}(Q)^{\otimes kr^t \times kr^t}$. Corollary 27 implies that

$$d_{\text{TV}}\left((M_{\text{R}})_{F_i'',F_j''}, \mathcal{N}(0,1)^{\otimes r\ell \times r\ell}\right) = O\left(r^{2t} \cdot (kr\ell)^{-3}\right)$$

for each $1 \le i, j \le k$. Since the submatrices $(M_{\text{R}})_{F_i'',F_j''}$ of M_{R} are independent, it follows that

$$d_{\text{TV}}\left(M_{\text{R}}, \mathcal{N}(0,1)^{\otimes kr\ell \times kr\ell}\right) = O\left(k^2 r^{2t} \cdot (kr\ell)^{-3}\right) = O\left((kr\ell)^{-1}\right)$$

by the tensorization property of total variation in Fact 15, completing the proof of the lemma. ■

The next lemma is immediate but makes explicit the precise guarantees for Step 4 of k-PDS-TO-ISBM.

Lemma 87 (Thresholding for ISBM) *Let F', F'', S and T be as in Lemma 86. Let $\mathcal{A}_4$ denote Step 4 of k-PDS-TO-ISBM with input M_{R} and output G'. Then*

$$\mathcal{A}_4\left(\frac{\mu(r-1)}{r} \cdot v_{S,F',F''}(K_{r,t}) v_{S,F',F''}(K_{r,t})^\top + \mathcal{N}(0,1)^{\otimes kr\ell \times kr\ell}\right) \sim \text{ISBM}_D(kr\ell, r, P_{11}, P_{12}, P_{22})$$
$$\mathcal{A}_4\left(\mathcal{N}(0,1)^{\otimes kr\ell \times kr\ell}\right) \sim \mathcal{G}(kr\ell, 1/2)$$

where P_{11}, P_{12} and P_{22} are as in Theorem 85.

Proof First observe that, since Lemma 29 implies that each column of $K_{r,t}$ contains exactly $(r-1)\ell$ entries equal to $1/\sqrt{r^t(r-1)}$ and ℓ entries equal to $(1-r)/\sqrt{r^t(r-1)}$, it follows that $v_{S,F',F''}(K_{r,t})$ contains $k(r-1)\ell$ entries equal to $1/\sqrt{r^t(r-1)}$ and $k\ell$ entries equal to $(1-r)/\sqrt{r^t(r-1)}$. Therefore there is a subset $T \subseteq [kr\ell]$ with $|T| = k\ell$ such that the $kr\ell \times kr\ell$ mean matrix $Z = v_{S,F',F''}(K_{r,t}) v_{S,F',F''}(K_{r,t})^\top$ has entries

$$Z_{ij} = \frac{1}{r^t(r-1)} \cdot \begin{cases} (r-1)^2 & \text{if } i,j \in S \\ -(r-1) & \text{if } i \in S \text{ and } j \notin S \text{ or } i \notin S \text{ and } j \in S \\ 1 & \text{if } i,j \notin S \end{cases}$$

Since the vertices of G' are randomly permuted, it follows by definition now that if

$$M_{\text{R}} \sim \mathcal{L}\left(\frac{\mu(r-1)}{r} \cdot v_{S,F',F''}(K_{r,t}) v_{S,F',F''}(K_{r,t})^\top + \mathcal{N}(0,1)^{\otimes kr\ell \times kr\ell}\right)$$

then $G' \sim \text{ISBM}_D(kr\ell, k\ell, P_{11}, P_{12}, P_{22})$, proving the first distributional equality in the lemma. The second distributional equality follows from the fact that $\Phi(0) = 1/2$. ■

We now complete the proof of Theorem 85 using a similar application of Lemma 16 as in the proof of Theorem 46.

Proof [Proof of Theorem 85] We apply Lemma 16 to the steps $\mathcal{A}_i$ of $\mathcal{A}$ under each of H_0 and H_1. Define the steps of $\mathcal{A}$ to map inputs to outputs as follows

$$(G,E) \xrightarrow{\mathcal{A}_1} (M_{\text{PD1}}, F) \xrightarrow{\mathcal{A}_2} (M_{\text{PD2}}, F') \xrightarrow{\mathcal{A}_3} (M_{\text{R}}, F'') \xrightarrow{\mathcal{A}_4} G'$$

Under H_1, consider Lemma 16 applied to the following sequence of distributions

$$\begin{aligned}
\mathcal{P}_0 &= \mathcal{G}_E(N,k,p,q) \\
\mathcal{P}_1 &= \mathcal{M}_{[m]\times[m]}(S\times S, \text{Bern}(p), \text{Bern}(Q)) \quad \text{where } S\sim\mathcal{U}_m(F) \\
\mathcal{P}_2 &= \mathcal{M}_{[kr^t]\times[kr^t]}(S\times S, \text{Bern}(p), \text{Bern}(Q)) \quad \text{where } S\sim\mathcal{U}_{kr^t}(F') \\
\mathcal{P}_3 &= \frac{\mu(r-1)}{r}\cdot v_{S,F',F''}(K_{r,t})v_{S,F',F''}(K_{r,t})^\top + \mathcal{N}(0,1)^{\otimes kr\ell\times kr\ell} \quad \text{where } S\sim\mathcal{U}_{kr^t}(F') \\
\mathcal{P}_4 &= \text{ISBM}_D(kr\ell, r, P_{11}, P_{12}, P_{22})
\end{aligned}$$

Applying Lemma 23, we can take

$$\epsilon_1 = 4k\cdot\exp\left(-\frac{Q^2N^2}{48pkm}\right) + \sqrt{\frac{C_Qk^2}{2m}}$$

where $C_Q = \max\left\{\frac{Q}{1-Q}, \frac{1-Q}{Q}\right\}$. The step $\mathcal{A}_2$ is exact and we can take $\epsilon_2 = 0$. Applying Lemma 86 and averaging over $S\sim\mathcal{U}_{kr^t}(F')$ using the conditioning property of total variation in Fact 15 yields that we can take $\epsilon_3 = O\left((kr\ell)^{-1}\right)$. By Lemma 87, Step 4 is exact and we can take $\epsilon_4 = 0$. By Lemma 16, we therefore have that

$$d_{\text{TV}}\left(\mathcal{A}\left(\mathcal{G}_E(N,k,p,q)\right), \text{ISBM}(n,r,P_{11},P_{12},P_{22})\right) = O\left(\frac{k}{\sqrt{N}} + e^{-\Omega(N^2/km)} + (kr\ell)^{-1}\right)$$

which proves the desired result in the case of H_1. Under H_0, consider the distributions

$$\begin{aligned}
\mathcal{P}_0 &= \mathcal{G}(N,q) \\
\mathcal{P}_1 &= \text{Bern}(Q)^{\otimes m\times m} \\
\mathcal{P}_2 &= \text{Bern}(Q)^{\otimes kr^t\times kr^t} \\
\mathcal{P}_3 &= \mathcal{N}(0,1)^{\otimes kr\ell\times kr\ell} \\
\mathcal{P}_4 &= \mathcal{G}(kr\ell, 1/2)
\end{aligned}$$

As above, Lemmas 23, 86 and 87 imply that we can take

$$\epsilon_1 = 4k\cdot\exp\left(-\frac{Q^2N^2}{48pkm}\right), \quad \epsilon_2 = 0, \quad \epsilon_3 = O\left((kr\ell)^{-1}\right) \quad \text{and} \quad \epsilon_4 = 0$$

By Lemma 16, we therefore have that

$$d_{\text{TV}}\left(\mathcal{A}\left(\mathcal{G}(N,q)\right), \mathcal{G}(n,1/2)\right) = O\left(e^{-\Omega(N^2/kn)} + (kr\ell)^{-1}\right)$$

which completes the proof of the theorem. ∎

We now prove that a slight modification to this reduction will map to all P_0 with $\min\{P_0, 1-P_0\} = \Omega(1)$ and to the setting where the density constraints in (12) hold exactly.

Corollary 88 (Reduction to Arbitrary P_0) *Let $0 < q < p \leq 1$ be constant and let N, r, k, E, ℓ and n be as in Theorem 85 with the additional condition that $kr^{3/2} = o(r^{2t})$. Suppose that P_0 satisfies $\min\{P_0, 1 - P_0\} = \Omega(1)$ and $\gamma \in (0, 1)$ satisfies that*

$$\gamma \leq \frac{c}{r^{t-1}\sqrt{\log(kr\ell)}}$$

for a sufficiently small constant $c > 0$. Then there is a $\mathrm{poly}(N)$ time reduction $\mathcal{A}$ from graphs on N vertices to graphs on n vertices satisfying that

$$\begin{aligned} &d_{TV}\left(\mathcal{A}\left(\mathcal{G}_E(N,k,p,q)\right), \textsc{isbm}_D\left(n, r, P_0 + \gamma, P_0 - \frac{\gamma}{k-1}, P_0 + \frac{\gamma}{(k-1)^2}\right)\right) \\ &\quad = O\left(\frac{k\mu^3 r^{3/2}}{r^{2t}} + \frac{k}{\sqrt{N}} + e^{-\Omega(N^2/km)} + (kr\ell)^{-1}\right) \\ &d_{TV}\left(\mathcal{A}\left(\mathcal{G}(N,q)\right), \mathcal{G}(n,P_0)\right) = O\left(e^{-\Omega(N^2/km)} + (kr\ell)^{-1}\right) \end{aligned}$$

Proof Consider the reduction $\mathcal{A}$ that adds a simple post-processing step to k-PDS-TO-ISBM as follows. On input graph G with N vertices:

1. Form the graph G_1 by applying k-PDS-TO-ISBM to G with parameters N, r, k, E, ℓ, n and μ where μ is given by

$$\mu = \frac{r^{t+1}}{(r-1)^2} \cdot \Phi^{-1}\left(\frac{1}{2} + \frac{1}{2} \cdot \min\{P_0, 1 - P_0\}^{-1} \cdot \gamma\right)$$

   and Φ^{-1} is the inverse of the standard normal CDF.

2. If $P_0 \leq 1/2$, output the graph G_2 formed by independently including each edge of G_1 in G_2 with probability $2P_0$. If $P_0 > 1/2$, form G_2 instead by including each edge of G_1 in G_2 and including each non-edge of G_1 in G_2 as an edge independently with probability $2P_0 - 1$.

This clearly runs in $\mathrm{poly}(N)$ time and it suffices to establish its approximate Markov transition properties. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ denote the two steps above with input-output pairs (G, G_1) and (G_1, G_2), respectively. Let $C \subseteq [n]$ be a fixed subset of size n/r and define

$$\begin{aligned} &P_{11} = \Phi\left(\frac{\mu(r-1)^2}{r^{t+1}}\right), \quad P_{12} = \Phi\left(-\frac{\mu(r-1)}{r^{t+1}}\right) \quad \text{and} \quad P_{22} = \Phi\left(\frac{\mu}{r^{t+1}}\right) \\ &P'_{11} = P_0 + \gamma, \quad P'_{12} = P_0 - \frac{\gamma}{r-1} \quad \text{and} \quad P'_{22} = P_0 + \frac{\gamma}{(r-1)^2} \end{aligned}$$

We will show that

$$d_{\mathrm{TV}}\left(\mathcal{A}_2\left(\textsc{isbm}_D\left(n, C, P_{11}, P_{12}, P_{22}\right)\right), \textsc{isbm}_D\left(n, C, P'_{11}, P'_{12}, P'_{22}\right)\right) = O\left(\frac{k\mu^3 r^{3/2}}{r^{2t}}\right) = o(1) \tag{13}$$

where the upper bound is $o(1)$ since $kr^{3/2} = o(r^{2t})$. First consider the case where $P_0 \leq 1/2$. Step 2 above yields by construction that

$$\mathcal{A}_2\left(\textsc{isbm}_D\left(n, C, P_{11}, P_{12}, P_{22}\right)\right) \sim \textsc{isbm}_D\left(n, C, 2P_0P_{11}, 2P_0P_{12}, 2P_0P_{22}\right)$$

Suppose that $X(r) \in \{0,1\}^m$ is sampled by first sampling $X' \sim \text{Bin}(m,r)$ and then letting X be selected uniformly at random from all elements of $\{0,1\}^m$ with support size X'. It follows that $X(r) \sim \text{Bern}(r)^{\otimes m}$ since both distributions are permutation-invariant and their support sizes have the same distribution. Now the data-processing inequality in Fact 15 implies that

$$d_{\text{TV}}\left(\text{Bern}(r)^{\otimes m}, \text{Bern}(r')^{\otimes m}\right) = d_{\text{TV}}\left(X(r), X(r')\right) \le d_{\text{TV}}\left(\text{Bin}(m,r), \text{Bin}(m,r')\right)$$

which can be upper bounded with Lemma 18. Using the fact that the edge indicators of ISBM conditioned on C are independent, the tensorization property in Fact 15 and Lemma 18, we now have that

$$\begin{aligned}
&d_{\text{TV}}\left(\text{ISBM}_D\left(n, C, 2P_0P_{11}, 2P_0P_{12}, 2P_0P_{22}\right), \text{ISBM}_D\left(n, C, P'_{11}, P'_{12}, P'_{22}\right)\right) \\
&\quad\le d_{\text{TV}}\left(\text{Bern}(2P_0P_{11})^{\otimes\binom{n/r}{2}}, \text{Bern}(P'_{11})^{\otimes\binom{n/r}{2}}\right) \\
&\qquad + d_{\text{TV}}\left(\text{Bern}(2P_0P_{12})^{\otimes\frac{n^2(r-1)}{r^2}}, \text{Bern}(P'_{12})^{\otimes\frac{n^2(r-1)}{r^2}}\right) \\
&\qquad + d_{\text{TV}}\left(\text{Bern}(2P_0P_{22})^{\otimes\binom{n(1-1/r)}{2}}, \text{Bern}(P'_{22})^{\otimes\binom{n(1-1/r)}{2}}\right) \\
&\quad\le \left|2P_0P_{11} - P'_{11}\right| \cdot \sqrt{\frac{\binom{n/r}{2}}{2P'_{11}(1-P'_{11})}} + \left|2P_0P_{12} - P'_{12}\right| \cdot \sqrt{\frac{n^2(r-1)}{2r^2P'_{12}(1-P'_{12})}} \\
&\qquad + \left|2P_0P_{22} - P'_{22}\right| \cdot \sqrt{\frac{\binom{n(1-1/r)}{2}}{2P'_{22}(1-P'_{22})}} \\
&\quad\le \left|2P_0P_{11} - P'_{11}\right| \cdot O\left(\frac{n}{r}\right) + \left|2P_0P_{12} - P'_{12}\right| \cdot O\left(\frac{n}{\sqrt{r}}\right) + \left|2P_0P_{22} - P'_{22}\right| \cdot O(n)
\end{aligned}$$

where the third inequality uses the fact that P'_{11}, P'_{12} and P'_{22} are each bounded away from 0 and 1. Observe that the definition of μ ensures

$$\frac{1}{2} + \frac{1}{2P_0} \cdot \gamma = \Phi\left(\frac{\mu(r-1)^2}{r^{t+1}}\right)$$

which implies that $2P_0P_{11} = P'_{11}$. We now use a standard Taylor approximation for the error function $\Phi(x) - 1/2$ around zero, given by $\Phi(x) = \frac{1}{2} + \frac{x}{\sqrt{2\pi}} + O(x^3)$ when $x \in (-1,1)$. Observe that

$$\begin{aligned}
\left|2P_0P_{12} - P'_{12}\right| &= 2P_0 \cdot \left|\Phi\left(-\frac{\mu(r-1)}{r^{t+1}}\right) - \frac{1}{2} + \frac{\gamma}{2P_0(r-1)}\right| \\
&= 2P_0 \cdot \left|\Phi\left(-\frac{\mu(r-1)}{r^{t+1}}\right) - \frac{1}{2} + \frac{1}{r-1}\left(\Phi\left(\frac{\mu(r-1)^2}{r^{t+1}}\right) - \frac{1}{2}\right)\right| \\
&= O\left(\frac{\mu^3 r^2}{r^{3t}}\right)
\end{aligned}$$

An analogous computation shows that $|2P_0P_{22} - P'_{22}| = O\left(\mu^3/r^{3t-1}\right)$. Combining all of these bounds now yields Equation (13) after noting that $n = kr\ell = O(kr^t)$ implies that $n\mu^3 r^{3/2}/r^{3t} =$

$O(kr^{3/2}/r^{2t})$. A nearly identical argument considering the complement of the graph G_1 and replacing with P_0 with $1 - P_0$ establishes Equation (13) in the case when $P_0 > 1/2$. Now observe that

$$\mathcal{A}_2\left(\mathcal{G}(n, 1/2)\right) \sim \mathcal{G}(n, P_0)$$

by definition. Now consider applying Lemma 16 to the steps $\mathcal{A}_1$ and $\mathcal{A}_2$ using an analogous recipe as in the proof of Theorem 85. We have that ϵ_1 is bounded by Theorem 85 and ϵ_2 is bounded by the argument above. Note that in order to apply Theorem 85 here, it must follow that the required bound on μ is met. Observe that

$$\gamma = 2P_0\left(\Phi\left(\frac{\mu(r-1)^2}{r^{t+1}}\right) - \frac{1}{2}\right) = \Theta\left(\frac{\mu}{r^{t-1}}\right)$$

and hence if γ satisfies the upper bound in the statement of the corollary for a sufficiently small constant c, then μ satisfies the requirement in Theorem 85 since p and q are constant. This application of Lemma 16 now yields the desired two approximate Markov transition properties and completes the proof of the corollary. ■

We now show that setting parameters in the reduction of Corollary 88 as in the recipe set out in Theorems 4 and 82 now shows that we can fill out the parameter space for ISBM obeying the edge density constraints of (12) below the Kesten-Stigum threshold. This proves the following computational lower bound for ISBM. We remark that typically the parameter regime of interest for the k-block stochastic block model is when $k = n^{o(1)}$, and thus the conditions (T) and $k = o(n^{1/3})$ are only mild restrictions here. Note that the condition (T) here is the same condition that was introduced in Section L.1.

Theorem 5 (Lower Bounds for ISBM) *Suppose that (n, k) satisfy condition (T), that k is prime or $k = \omega_n(1)$ and $k = o(n^{1/3})$, and suppose that $P_0 \in (0, 1)$ satisfies $\min\{P_0, 1 - P_0\} = \Omega_n(1)$. Consider the testing problem $\mathrm{ISBM}(n, k, P_{11}, P_{12}, P_{22})$ where*

$$P_{11} = P_0 + \gamma, \quad P_{12} = P_0 - \frac{\gamma}{k-1} \quad \text{and} \quad P_{22} = P_0 + \frac{\gamma}{(k-1)^2}$$

Then the k-PC conjecture or k-PDS conjecture for constant $0 < q < p \le 1$ both imply that there is a computational lower bound for $\mathrm{ISBM}(n, k, P_{11}, P_{12}, P_{22})$ at all levels of signal below the Kesten-Stigum threshold of $\gamma^2 = \tilde{o}(k^2/n)$.

Proof It suffices to show that the reduction $\mathcal{A}$ in Corollary 88 applied with $r \ge 2$ fills out all of the possible growth rates specified by the computational lower bound $\gamma^2 = \tilde{o}(k^2/n)$ and the other conditions in the theorem statement. Fix a constant pair of probabilities $0 < q < p \le 1$ and any sequence of parameters (n, k, γ, P_0) all of which are implicitly functions of n such that (n, k) satisfies (T) and

$$\gamma^2 \le \frac{k^2}{w' \cdot n \log n}, \quad 2(w')^2 k \le n^{1/3} \quad \text{and} \quad \min\{P_0, 1 - P_0\} = \Omega_n(1)$$

for sufficiently large n and $w' = w'(n) = (\log n)^c$ for a sufficiently large constant $c > 0$. Now let $w = w(n) \to \infty$ be an arbitrarily slow-growing increasing positive integer-valued function at least satisfying that $w(n) = n^{o(1)}$. As in the proof of Theorem 4, we now specify the following in order to fulfill the criteria in Condition E.1:

1. a sequence (N, k_N) such that the k-PDS(N, k_N, p, q) is hard according to Conjecture 3; and
2. a sequence (n', k', γ, P_0) with a subsequence that satisfies three conditions: (2.1) the parameters on the subsequence are in the regime of the desired computational lower bound for ISBM; (2.2) they have the same growth rate as (n, k, γ, P_0) on this subsequence; and (2.3) such that ISBM with the parameters on this subsequence can be produced by $\mathcal{A}$ with input k-PDS(N, k_N, p, q).

As discussed in Section E.2, this is sufficient to prove the theorem. We choose these parameters as follows:

- let $k' = r$ be the smallest prime satisfying that $k \le r \le 2k$, which exists by Bertrand's postulate and can be found in poly(n) time;
- let t be such that r^t is the closest power of r to $\sqrt{n}$ and let

$$k_N = \left\lfloor \frac{1}{2}\left(1 + \frac{p}{Q}\right)^{-1} w^{-2} \cdot \min\left\{r^t, \sqrt{n}\right\} \right\rfloor$$

  where $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}}\left(\sqrt{q} - 1\right)$; and
- let $n' = k_N r \ell$ where $\ell = \frac{r^t - 1}{r-1}$ and let $N = w k_N^2$.

Note that we have that $w^2 r \le n^{1/3}$ since $r \le 2k$. Now observe that we have the following bounds

$$\begin{aligned}
n' &\asymp k_N r^t \asymp \left(w^{-2} \cdot \min\left\{\frac{r^t}{\sqrt{n}}, 1\right\} \cdot \frac{r^t}{\sqrt{n}}\right) n \\
k_N r^{3/2} &\lesssim w^{-2} \cdot \min\left\{r^t, \sqrt{n}\right\} \cdot w^{-3}\sqrt{n} \lesssim \left(w^{-4} \cdot \frac{n}{r^{2t}}\right) r^{2t} \\
m &\le 2\left(\frac{p}{Q} + 1\right) w k_N^2 \le \left(w^{-3} \cdot \frac{\sqrt{n}}{r^t}\right) k_N r^t \\
k_N r \ell &\le \text{poly}(N) \\
\gamma^2 &\le \frac{k^2}{w' \cdot n \log n} = \frac{1}{w' \cdot r^{2t-2} \log(k_N r \ell)} \cdot \frac{r^{2t} \log(k_N r \ell)}{n \log n} \\
\gamma^2 &\lesssim \frac{r^2}{w' \cdot n' \log n'} \left(w^{-2} \cdot \min\left\{\frac{r^t}{\sqrt{n}}, 1\right\} \cdot \frac{r^t}{\sqrt{n}}\right) \cdot \frac{\log n'}{\log n} \lesssim \frac{r^2}{w' \cdot w^2 \cdot n' \log n'} \cdot \frac{r^t}{\sqrt{n}}
\end{aligned}$$

where m is the smallest multiple of k_N larger $\left(\frac{p}{Q} + 1\right) N$. Now observe that as long as $\sqrt{n} = \tilde{\Theta}(r^t)$ then: (2.1) the last inequality above on γ^2 would imply that (n', k', γ, P_0) is in the desired hard regime; (2.2) n and n' have the same growth rate since $w = n^{o(1)}$, and k and $k' = r$ have the same growth rate since either $k' = k$ or $k' = \Theta(k) = \omega(1)$; and (2.3) the middle four bounds above imply that taking c large enough yields the conditions needed to apply Corollary 88 to yield the desired reduction. By Lemma 80, there is an infinite subsequence of the input parameters such that $\sqrt{n} = \tilde{\Theta}(r^t)$, which concludes the proof as in Theorem 4. ■

M.2. Testing Hidden Partition Models

In this section, we establish statistical-computational gaps based on the k-PC and k-PDS conjectures for detection in the Gaussian and bipartite hidden partition models introduced in Sections B.5 and E.3. These two models are bipartite analogues of the subgraph variants of the k-block stochastic block model in the constant edge density regime. Specifically, they are multiple-community variants of the subgraph stochastic block model considered in Brennan et al. (2018).

The motivation for considering these two models is to illustrate the versatility of Bernoulli rotations as a reduction primitive. These two models are structurally very different from planted clique yet can be produced through Bernoulli rotations for appropriate choices of the output mean vectors $A_1, A_2, \ldots, A_m$. The mean vectors specified in the reduction are vectorizations of the slices of the design tensor $T_{r,t}$ constructed based on the incidence geometry of $\mathbb{F}_r^t$. The definition of $T_{r,t}$ and several of its properties can be found in Section G.3. The reduction in this section demonstrates that natural applications of Bernoulli rotations can require more involved constructions than $K_{r,t}$ in order to produce tight computational lower bounds.

We begin by reviewing the definitions of the two main models considered in this section – Gaussian and bipartite hidden partition models – which were introduced in Sections B.5 and E.3.

Definition 89 (Gaussian Hidden Partition Models) *Let n, r and K be positive integers, let $\gamma \in \mathbb{R}$ and let $C = (C_1, C_2, \ldots, C_r)$ be a sequence of disjoint K-subsets of $[n]$. Let $D = (D_1, D_2, \ldots, D_r)$ be another such sequence. The distribution $\text{GHPM}_D(n, r, C, D, \gamma)$ over matrices $M \in \mathbb{R}^{n \times n}$ is such that $M_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(\mu_{ij}, 1)$ where*

$$\mu_{ij} = \begin{cases} \gamma & \text{if } i \in C_h \text{ and } j \in D_h \text{ for some } h \in [r] \\ -\frac{\gamma}{r-1} & \text{if } i \in C_{h_1} \text{ and } j \in D_{h_2} \text{ where } h_1 \neq h_2 \\ 0 & \text{otherwise} \end{cases}$$

for each $i, j \in [n]$. Furthermore, let $\text{GHPM}_D(n, r, K, \gamma)$ denote the mixture over $\text{GHPM}_D(n, r, C, D, \gamma)$ induced by choosing C and D independently and uniformly at random.

Definition 90 (Bipartite Hidden Partition Models) *Let n, r, K, C and D be as in Definition 89 and let $P_0, \gamma \in (0, 1)$ be such that $\gamma/r \leq P_0 \leq 1 - \gamma$. The distribution $\text{BHPM}_D(n, r, C, D, P_0, \gamma)$ over bipartite graphs G with two parts of size n, each indexed by $[n]$, such that each edge (i, j) is included in G independently with the following probabilities*

$$\mathbb{P}\left[(i, j) \in E(G)\right] = \begin{cases} P_0 + \gamma & \text{if } i \in C_h \text{ and } j \in D_h \text{ for some } h \in [r] \\ P_0 - \frac{\gamma}{r-1} & \text{if } i \in C_{h_1} \text{ and } j \in D_{h_2} \text{ where } h_1 \neq h_2 \\ P_0 & \text{otherwise} \end{cases}$$

for each $i, j \in [n]$. Let $\text{BHPM}_D(n, r, K, P_0, \gamma)$ denote the mixture over $\text{BHPM}_D(n, r, C, D, P_0, \gamma)$ induced by choosing C and D independently and uniformly at random.

The problems we consider in this section are the two simple hypothesis testing problems GHPM and BHPM from Section E.3, given by

$$\begin{aligned} H_0 &: M \sim \mathcal{N}(0, 1)^{\otimes n \times n} \quad \text{and} \quad H_1 : M \sim \text{GHPM}(n, r, K, \gamma) \\ H_0 &: G \sim \mathcal{G}_B(n, n, P_0) \quad \text{and} \quad H_1 : G \sim \text{BHPM}(n, r, K, P_0, \gamma) \end{aligned}$$

An important remark is that the hypothesis testing formulations above for these two problems seem to have different computational and statistical barriers from the tasks of recovering C and D. We now state the following lemma, giving guarantees for a natural polynomial-time test and exponential time test for GHPM. The proof of this lemma is tangential to the main focus of this section – computational lower bounds for GHPM and BHPM – and is deferred to Appendix R.2.

Lemma 91 (Tests for GHPM) *Given a matrix $M \in \mathbb{R}^{n\times n}$, let $s_C(M) = \sum_{i,j=1}^n M_{ij}^2 - n^2$ and*

$$s_I(M) = \max_{C,D} \left\{ \sum_{h=1}^{r} \sum_{i\in C_h} \sum_{j\in D_h} M_{ij} \right\}$$

where the maximum is over all pairs (C, D) of sequences of disjoint K-subsets of $[n]$. Let $w = w(n)$ be any increasing function with $w(n) \to \infty$ as $n \to \infty$. We prove the following:

1. *If $M \sim \text{GHPM}_D(n, r, K, \gamma)$, then with probability $1 - o_n(1)$ it holds that*

$$s_C(M) \geq rK^2\gamma^2 + \frac{rK^2}{r-1}\cdot\gamma^2 - w\left(n + \gamma K\sqrt{r} + \frac{K\gamma}{r}\right) \quad \text{and} \quad s_I(M) \geq rK^2\gamma - wr^{1/2}K$$

2. *If $M \sim \mathcal{N}(0,1)^{\otimes n\times n}$, then with probability $1 - o_n(1)$ it holds that*

$$s_C(M) \leq wn \quad \text{and} \quad s_I(M) \leq 2rK^{3/2}w\sqrt{(\log n + \log r)}$$

This lemma implies upper bounds on the computational and statistical barriers for GHPM. Specifically, it implies that the variance test s_C succeeds above $\gamma_{\text{comp}}^2 = \tilde{\Theta}(n/rK^2)$ and the search test s_I succeeds above $\gamma_{\text{IT}}^2 = \tilde{\Theta}(1/K)$. Thus, showing that there is a computational barrier at this level of signal γ_{comp} is sufficient to show that there is a nontrivial statistical-computational gap for GHPM. For P_0 with $\min\{P_0, 1 - P_0\} = \Omega(1)$, analogous tests show the same upper bounds on γ_{comp} and γ_{IT} for BHPM.

Consider the case when $n = rK$, which corresponds to a testing variant of the bipartite k-block stochastic block model. In this case, the upper bounds shown by the previous lemma coincide at $\gamma_{\text{comp}}^2, \gamma_{\text{IT}}^2 = O(r/n)$ and hence do not support the existence of a statistical-computational gap. The subgraph formulation in which $rK \ll n$ seems crucial to yielding a testing problem with a statistical-computational gap. We also remark that while this testing formulation when $n = rK$ may not have a gap, the task of recovering C and D likely shares the gap conjectured in the k-block stochastic block model. Specifically, the conjectured computational barrier at the Kesten-Stigum threshold lies at $\gamma^2 = \tilde{\Theta}(r^2/n)$, which lies well above the r/n limit in the testing formulation.

The rest of this section is devoted to giving our main reduction k-PDS-TO-GHPM showing a computational barrier at $\gamma^2 = \tilde{o}(n/rK^2)$. This reduction is shown in Figure 13 and its approximate Markov transition guarantees are stated in the theorem below. The intuition behind why our reduction is tight to the algorithm s_C is as follows. Bernoulli rotations are approximately ℓ_2-norm preserving in the signal to noise ratio if the output dimension is comparable to the input dimension with $m \asymp n$. Much of the effort in constructing $T_{r,t}$ and $M_{r,t}$ in Section G.3 was devoted to the linear functions L which are crucial in designing $M_{r,t}$ to be nearly square and hence achieve $m \asymp n$ in Bernoulli rotations. Any reduction that is approximately ℓ_2-norm preserving in the signal to noise ratio will be tight to a variance test such as s_C.

Algorithm k-PDS-TO-GHPM

Inputs: k-PDS instance $G \in \mathcal{G}_N$ with dense subgraph size k that divides N, and the following parameters

- partition E, edge probabilities $0 < q < p \le 1$, $Q \in (0,1)$ and m as in Figure 12
- refinement parameter s and number of vertices $n = ksr^t$ where r is a prime number, $\ell = \frac{r^t-1}{r-1}$ for some $t \in \mathbb{N}$ satisfy that $m \le ks(r-1)\ell \le \text{poly}(N)$
- mean parameter $\mu \in (0,1)$ as in Figure 12

1. *Symmetrize and Plant Diagonals*: Compute $M_{\text{PD1}} \in \{0,1\}^{m \times m}$ and F as in Step 1 of Figure 12.
2. *Pad and Further Partition*: Form M_{PD2} and F' as in Step 2 of Figure 12 modified so that M_{PD2} is a $ks(r-1)\ell \times ks(r-1)\ell$ matrix and each F'_i has size $s(r-1)\ell$. Let F^s be the partition of $[ks(r-1)\ell]$ into ks parts of size $(r-1)\ell$ by refining F' by splitting each of its parts into s parts of equal size arbitrarily.
3. *Bernoulli Rotations*: Let F^o be a partition of $[ksr^t]$ into ks equally sized parts. Now compute the matrix $M_{\text{R}} \in \mathbb{R}^{ksr^t \times ksr^t}$ as follows:
   (1) For each $i, j \in [ks]$, flatten the $(r-1)\ell \times (r-1)\ell$ submatrix $(M_{\text{P}})_{F^s_i, F^s_j}$ into a vector $V_{ij} \in \mathbb{R}^{(r-1)^2\ell^2}$ and let $A = M_{r,t}^\top \in \mathbb{R}^{r^{2t} \times (r-1)^2\ell^2}$ as in Definition 34.
   (2) Apply BERN-ROTATIONS to V_{ij} with matrix A, rejection kernel parameter $R_{\text{RK}} = ksr^t$, Bernoulli probabilities $0 < Q < p \le 1$, output dimension r^{2t}, $\lambda = \sqrt{1 + (r-1)^{-1}}$ and mean parameter μ.
   (3) Set the entries of $(M_{\text{R}})_{F^o_i, F^o_j}$ to be the entries of the output in (2) unflattened into a matrix.
4. *Permute and Output*: Output the matrix M_{R} with its rows and columns independently permuted uniformly at random.

Figure 13: Reduction from k-partite planted dense subgraph to gaussian hidden partition models.

The key to the reduction k-PDS-TO-GHPM lies in the construction of $T_{r,t}$ and $M_{r,t}$ in Section G.3. The rest of the proof of the following theorem is similar to the proofs in the previous section. We omit details that are similar for brevity. We recall from Section E.4 that, given a matrix $M \in \mathbb{R}^{n \times n}$, the matrix $M_{S,T} \in \mathbb{R}^{k \times k}$ where S, T are k-subsets of $[n]$ refers to the minor of M restricted to the row indices in S and column indices in T. Furthermore, $(M_{S,T})_{i,j} = M_{\sigma_S(i), \sigma_T(j)}$ where $\sigma_S : [k] \to S$ is the unique order-preserving bijection and σ_T is analogously defined.

Theorem 92 (Reduction to GHPM) *Let N be a parameter and $r = r(N) \ge 2$ be a prime number. Fix initial and target parameters as follows:*

- Initial k-BPDS Parameters: k, N, p, q *and* E *as in Theorem 85.*
- Target GHPM Parameters: (n, r, K, γ) *where* $n = ksr^t$, $K = kr^{t-1}$ *and* $\ell = \frac{r^t-1}{r-1}$ *for some parameters* $t = t(N), s = s(N) \in \mathbb{N}$ *satisfying that that*

$$m \leq ks(r-1)\ell \leq \text{poly}(N)$$

where m *and* Q *are as in Theorem 92. The target level of signal* γ *is given by* $\gamma = \frac{\mu(r-1)}{r^t\sqrt{r}}$ *where*

$$\mu \leq \frac{1}{2\sqrt{6\log(ksr^t) + 2\log(p-Q)^{-1}}} \cdot \min\left\{\log\left(\frac{p}{Q}\right), \log\left(\frac{1-Q}{1-p}\right)\right\}$$

Let $\mathcal{A}(G)$ *denote* k-PDS-TO-GHPM *applied to the graph* G *with these parameters. Then* $\mathcal{A}$ *runs in* $\text{poly}(N)$ *time and it follows that*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}_E(N,k,p,q)\right), \text{GHPM}_D(n,r,K,\gamma)\right) = O\left(\frac{k}{\sqrt{N}} + e^{-\Omega(N^2/km)} + (ksr^t)^{-1}\right)$$
$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}(N,q)\right), \mathcal{N}(0,1)^{\otimes n\times n}\right) = O\left(e^{-\Omega(N^2/km)} + (ksr^t)^{-1}\right)$$

In order to state the approximate Markov transition guarantees of the Bernoulli rotations step of k-PDS-TO-GHPM, we need the formalism from Section G.3 to describe the matrix $M_{r,t}$, tensor $T_{r,t}$ and their community alignment properties. While this will require a plethora of cumbersome notation, the goal of the ensuing discussion is simple – we will show that Lemma 36 guarantees that stitching together the individual applications of BERN-ROTATIONS in Step 3 of k-PDS-TO-GHPM yields a valid instance of GHPM.

Recall $\mathcal{C}(M^{1,1}, M^{1,2}, \ldots, M^{ks,ks})$ denotes the concatenation of k^2s^2 matrices $M^{i,j} \in \mathbb{R}^{r^t\times r^t}$ into a $ksr^t \times ksr^t$ matrix, as introduced in Section G.3. Given a partition F of $[ksr^t]$ into ks equally sized parts, let $\mathcal{C}_F(M^{1,1}, M^{1,2}, \ldots, M^{ks,ks})$ denote the concatenation of the $M^{i,j}$, where now the entries of $M^{i,j}$ appear in $\mathcal{C}_F$ on the index set $F_i \times F_j$. For consistency, we fix a canonical embedding of the row and column indices of $\mathbb{R}^{r^t\times r^t}$ to $F_i \times F_j$ by always preserving the order of indices.

Let F^o and F^s be fixed partitions of $[ksr^t]$ and $[ks(r-1)\ell]$ into k parts of size r^t and $(r-1)\ell$, respectively, and let $S \subseteq [ks(r-1)\ell]$ be such that $|S| = k$ and S intersects each part of F^s in at most one element. Now let $\mathbf{M}_{S,F^s,F^o}(T_{r,t}) \in \mathbb{R}^{ksr^t\times ksr^t}$ be the matrix

$$\mathbf{M}_{S,F^s,F^o}(T_{r,t}) = \mathcal{C}_{F^o}\left(M^{1,1}, M^{1,2}, \ldots, M^{ks,ks}\right) \quad \text{where} \quad M^{i,j} = \begin{cases} T_{r,t}^{(V_{t_i}, V_{t_j}, L_{ij})} & \text{if } S \cap F_i^s \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

where t_i, t_j and L_{ij} are given by:

- let $\sigma : [ks(r-1)\ell] \to [ks(r-1)\ell]$ be the unique bijection transforming the partition F^s to the canonical contiguous partition $\{1, \ldots, (r-1)\ell\} \cup \cdots \cup \{(ks-1)(r-1)\ell+1, \ldots, ks(r-1)\ell\}$ while preserving ordering on each part F_i^s for $1 \leq i \leq ks$;
- let s_i' be the unique element in $\sigma(S \cap F_i^s)$ for each i for which this intersection is nonempty, and let s_i be the unique positive integer with $1 \leq s_i \leq (r-1)\ell$ and $s_i \equiv s_i' \pmod{(r-1)\ell}$; and

- t_i, t_j and L_{ij} are as in Lemma 36 given these s_i i.e. t_i and t_j are the unique $1 \le t_i, t_j \le \ell$ such that $t_i \equiv s_i \pmod{\ell}$ and $t_j \equiv s_j \pmod{\ell}$ and $L_{ij} : \mathbb{F}_r \to \mathbb{F}_r$ is given by $L_{ij}(x) = a_i x + a_j$ where $a_i = \lceil s_i/\ell \rceil$ and $a_j = \lceil s_j/\ell \rceil$.

The next lemma makes explicit the implications of Lemma 26 and Lemma 36 for the approximate Markov transition guarantees of Step 3 in k-PDS-TO-GHPM. The proof follows a similar structure to the proof of Lemma 86 and we omit identical details.

Lemma 93 (Bernoulli Rotations for GHPM) *Let F^o and F^s be a fixed partitions of $[ksr^t]$ and $[ks(r-1)\ell]$ into k parts of size r^t and $(r-1)\ell$, respectively, and let $S \subseteq [ksr^t]$ be such that $|S| = k$ and $|S \cap F_i^s| \le 1$ for each $1 \le i \le ks$. Let $\mathcal{A}_3$ denote Step 3 of k-PDS-TO-GHPM with input M_{PD2} and output M_{R}. Suppose that p, Q and μ are as in Theorem 85, then it follows that*

$$d_{TV}\Big(\mathcal{A}_3\left(\mathcal{M}_{[ks(r-1)\ell]\times[ks(r-1)\ell]}\left(S\times S, \text{Bern}(p), \text{Bern}(Q)\right)\right), \\ \mathcal{L}\left(\mu\sqrt{\frac{r-1}{r}}\cdot \mathbf{M}_{S,F^s,F^o}(T_{r,t}) + \mathcal{N}(0,1)^{\otimes ksr^t \times ksr^t}\right)\Big) = O\left((ksr^t)^{-1}\right)$$

$$d_{TV}\left(\mathcal{A}_3\left(\text{Bern}(Q)^{\otimes ks(r-1)\ell \times ks(r-1)\ell}\right), \mathcal{N}(0,1)^{\otimes ksr^t \times ksr^t}\right) = O\left((ksr^t)^{-1}\right)$$

and furthermore, for all such subsets S, it holds that the matrix $\mathbf{M}_{S,F^s,F^o}(T_{r,t})$ has zero entries other than in a $kr^t \times kr^t$ submatrix, which is also r-block as defined in Section G.3.

Proof Define s_i', s_i, t_i and L_{ij} as in the preceding discussion for all i, j with $S \cap F_i^s$ and $S \cap F_j^s$ nonempty. Let (1) and (2) denote the following two cases:

1. $M_{\text{PD2}} \sim \mathcal{M}_{[ks(r-1)\ell]\times[ks(r-1)\ell]}\left(S\times S, \text{Bern}(p), \text{Bern}(Q)\right)$; and
2. $M_{\text{PD2}} \sim \text{Bern}(Q)^{\otimes ks(r-1)\ell \times ks(r-1)\ell}$.

Now define the matrix M_{R}' with independent entries such that

$$\left(M_{\text{R}}'\right)_{F_i^s, F_j^s} \sim \begin{cases} \mu\sqrt{\frac{r-1}{r}}\cdot T_{r,t}^{(V_{t_i}, V_{t_j}, L_{ij})} + \mathcal{N}(0,1)^{\otimes r^t \times r^t} & \text{if (1) holds, } S\cap F_i^s \neq \emptyset \text{ and } S \cap F_j^s \neq \emptyset \\ \mathcal{N}(0,1)^{\otimes r^t \times r^t} & \text{otherwise if (1) holds or if (2) holds} \end{cases}$$

for each $1 \le i, j \le ks$. The vectorization and ordering conventions we adopt imply that if $S \cap F_i^s \neq \emptyset$ and $S \cap F_j^s \neq \emptyset$, then the unflattening of the row with index $(s_i - 1)(r-1)\ell + s_j$ in $M_{r,t}$ is the approximate output mean of $\mathcal{A}_3$ on the minor $(M_{\text{R}})_{F_i^s, F_j^s}$ when applying Lemma 26 under (1). By Definition 34 and the definitions of a_i, t_i and L_{ij}, this unflattened row is exactly the matrix

$$M^{i,j} = T_{r,t}^{(V_{t_i}, V_{t_j}, L_{ij})}$$

Combining this observation with Lemmas 26 and 35 yields that under both (1) and (2), we have that

$$d_{\text{TV}}\left(\left(M_{\text{R}}\right)_{F_i^s, F_j^s}, \left(M_{\text{R}}'\right)_{F_i^s, F_j^s}\right) = O\left(r^{2t}\cdot (ksr^t)^{-3}\right)$$

for all $1 \le i, j \le ks$. Through the same argument as in Lemma 86, the tensorization property of total variation in Fact 15 now yields that $d_{\mathrm{TV}}\left(\mathcal{L}(M_{\mathrm{R}}), \mathcal{L}(M'_{\mathrm{R}})\right) = O\left((ksr^t)^{-1}\right)$ under both (1) and (2). Now note that the definition of $\mathcal{C}_{F^o}$ implies that

$$M'_{\mathrm{R}} \sim \begin{cases} \mu\sqrt{\frac{r-1}{r}} \cdot \mathbf{M}_{S,F^s,F^o}(T_{r,t}) + \mathcal{N}(0,1)^{\otimes ksr^t \times ksr^t} & \text{if (1) holds} \\ \mathcal{N}(0,1)^{\otimes ksr^t \times ksr^t} & \text{if (2) holds} \end{cases}$$

which completes the proof of the approximate Markov transition guarantees in the lemma statement. Now note that $\mathbf{M}_{S,F^s,F^o}(T_{r,t})$ is zero everywhere other than on the union U of the F_i^o over the i such that $S \cap F_i^s \neq \emptyset$. There are exactly k such i and thus $|U| = kr^t$. Note that r-block matrices remain r-block matrices under permutations of column and row indices, and therefore Lemma 36 implies the same conclusion if $\mathcal{C}$ is replaced by $\mathcal{C}_{F^o}$. Applying Lemma 36 to the submatrix of $\mathbf{M}_{S,F^s,F^o}(T_{r,t})$ restricted to the indices of U now completes the proof of the lemma. ■

We now complete the proof of Theorem 92, again applying Lemma 16 as in the proofs of Theorems 46 and 85. In this theorem, we let $\mathcal{U}_n^k(F)$ denote the uniform distribution over subsets $S \subseteq [n]$ of size k intersecting each part of the partition F in at most one element. When F has exactly k parts, this definition recovers the previously defined distribution $\mathcal{U}_n(F)$.

Proof [Proof of Theorem 92] Let the steps of $\mathcal{A}$ to map inputs to outputs as follows

$$(G, E) \xrightarrow{\mathcal{A}_1} (M_{\mathrm{PD1}}, F) \xrightarrow{\mathcal{A}_2} (M_{\mathrm{PD2}}, F^s) \xrightarrow{\mathcal{A}_3} (M_{\mathrm{R}}, F^o) \xrightarrow{\mathcal{A}_4} M'_{\mathrm{R}}$$

where here M'_{R} denotes the permuted form of M_{R} after Step 4. Under H_1, consider Lemma 16 applied to the following sequence of distributions

$$\begin{aligned}
\mathcal{P}_0 &= \mathcal{G}_E(N, k, p, q) \\
\mathcal{P}_1 &= \mathcal{M}_{[m]\times[m]}(S \times S, \mathrm{Bern}(p), \mathrm{Bern}(Q)) \quad \text{where } S \sim \mathcal{U}_m(F) \\
\mathcal{P}_2 &= \mathcal{M}_{[ks(r-1)\ell]\times[ks(r-1)\ell]}(S \times S, \mathrm{Bern}(p), \mathrm{Bern}(Q)) \quad \text{where } S \sim \mathcal{U}^k_{ks(r-1)\ell}(F^s) \\
\mathcal{P}_3 &= \mu\sqrt{\frac{r-1}{r}} \cdot \mathbf{M}_{S,F^s,F^o}(T_{r,t}) + \mathcal{N}(0,1)^{\otimes ksr^t \times ksr^t} \quad \text{where } S \sim \mathcal{U}^k_{ks(r-1)\ell}(F^s) \\
\mathcal{P}_4 &= \mathrm{GHPM}_D\left(ksr^t, r, kr^{t-1}, \frac{\mu(r-1)}{r^t\sqrt{r}}\right)
\end{aligned}$$

Let $C_Q = \max\left\{\frac{Q}{1-Q}, \frac{1-Q}{Q}\right\}$ and consider setting

$$\epsilon_1 = 4k \cdot \exp\left(-\frac{Q^2N^2}{48pkm}\right) + \sqrt{\frac{C_Q k^2}{2m}}, \quad \epsilon_2 = 0, \quad \epsilon_3 = O\left((ksr^t)^{-1}\right) \quad \text{and} \quad \epsilon_4 = 0$$

As in the proof of Theorem 85, Lemma 23 implies this is a valid choice of ϵ_1 and $\mathcal{A}_2$ is exact so we can take $\epsilon_2 = 0$. The choice of ϵ_3 is valid by applying Lemma 93 and averaging over $S \sim \mathcal{U}^k_{ks(r-1)\ell}(F^s)$ using the conditioning property of total variation in Fact 15. Now note that the $kr^t \times kr^t$ r-block submatrix of $\mathbf{M}_{S,F^s,F^o}(T_{r,t})$ has entries $\frac{r-1}{r^t\sqrt{r-1}}$ and $-\frac{1}{r^t\sqrt{r-1}}$. Thus the matrix $\mu\sqrt{\frac{r-1}{r}} \cdot \mathbf{M}_{S,F^s,F^o}(T_{r,t})$ is of the form of the mean matrix $(\mu_{ij})_{1\le i,j\le ksr^t}$ in Definition 89 for some choice of C and D where $K = kr^{t-1}$ and

$$\gamma = \mu\sqrt{\frac{r-1}{r}} \cdot \frac{r-1}{r^t\sqrt{r-1}} = \frac{\mu(r-1)}{r^t\sqrt{r}}$$

This implies that permuting the rows and columns of $\mathcal{P}_3$ yields $\mathcal{P}_4$ exactly with $\epsilon_4 = 0$. Applying Lemma 16 now yields the first bound in the theorem statement. Under H_0, consider the distributions

$$\mathcal{P}_0 = \mathcal{G}(N, q), \quad \mathcal{P}_1 = \text{Bern}(Q)^{\otimes m \times m}, \quad \mathcal{P}_2 = \text{Bern}(Q)^{\otimes ks(r-1)\ell \times ks(r-1)\ell},$$

$$\mathcal{P}_3 = \mathcal{P}_4 = \mathcal{N}(0,1)^{\otimes ksr^t \times ksr^t}$$

As above, Lemmas 23 and 93 imply that we can take $\epsilon_1 = 4k \cdot \exp\left(-\frac{Q^2N^2}{48pkm}\right)$ and ϵ_2, ϵ_3 and ϵ_4 as above. Lemma 16 now yields the second bound in the theorem statement. ■

We now append a final post-processing step to the reduction k-PDS-TO-GHPM to map to BHPM. The proof of the following corollary is similar to that of Corollary 88 and is deferred to Appendix R.2.

Corollary 94 (Reduction from GHPM to BHPM) *Let $0 < q < p \leq 1$ be constant and let the parameters k, N, E, r, ℓ, n, s and K be as in Theorem 92 with the additional condition that $k\sqrt{r} = o(r^{2t})$. Let $\gamma \in (0,1)$ be such that*

$$\gamma \leq \frac{c(r-1)}{r^t\sqrt{r\log(ksr^t)}}$$

for a sufficiently small constant $c > 0$. Suppose that P_0 satisfies $\min\{P_0, 1-P_0\} = \Omega(1)$. Then there is a $\text{poly}(N)$ time reduction $\mathcal{A}$ from graphs on N vertices to graphs on n vertices satisfying that

$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}_E(N,k,p,q)\right), \text{BHPM}_D(n,r,K,P_0,\gamma)\right) = O\left(\frac{k\mu^3\sqrt{r}}{r^{2t}} + \frac{k}{\sqrt{N}} + e^{-\Omega(N^2/km)} + (ksr^t)^{-1}\right)$$

$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}(N,q)\right), \mathcal{G}_B(N,N,P_0)\right) = O\left(e^{-\Omega(N^2/km)} + (ksr^t)^{-1}\right)$$

Collecting the results of this section, we arrive at the following computational lower bounds for GHPM and BHPM matching the efficient test s_C in Lemma 91.

Theorem 6 (Lower Bounds for GHPM and BHPM) *Suppose that $r^2K^2 = \tilde{\omega}(n)$ and $(\lceil r^2K^2/n \rceil, r)$ satisfies condition* (T)*, suppose r is prime or $r = \omega_n(1)$ and suppose that $P_0 \in (0,1)$ satisfies $\min\{P_0, 1-P_0\} = \Omega_n(1)$. Then the k-PC conjecture or k-PDS conjecture for constant $0 < q < p \leq 1$ both imply that there is a computational lower bound for each of $\text{GHPM}(n,r,K,\gamma)$ for all levels of signal $\gamma^2 = \tilde{o}(n/rK^2)$. This same lower bound also holds for $\text{BHPM}(n,r,K,P_0,\gamma)$ given the additional condition $n = o(rK^{4/3})$.*

Proof The proof of this theorem will follow that of Theorem 5 with several modifications. We begin by showing a lower bound for GHPM. It suffices to show that the reduction k-PDS-TO-GHPM fills out all of the possible growth rates specified by the computational lower bound $\gamma^2 = \tilde{o}(n/rK^2)$ and the other conditions in the theorem statement. Fix a constant pair of probabilities $0 < q < p \leq 1$ and any sequence of parameters (n, r, K, γ) all of which are implicitly functions of n such that $(\lceil r^2K^2/n \rceil, r)$ satisfies (T) and

$$\gamma^2 \leq c \cdot \frac{n}{w' \cdot rK^2 \log n} \quad \text{and} \quad r^2K^2 \geq w'n$$

for sufficiently large n and $w' = w'(n) = (\log n)^c$ for a sufficiently large constant $c > 0$. Now let $w = w(n) \to \infty$ be an arbitrarily slow-growing increasing positive integer-valued function at least satisfying that $w(n) = n^{o(1)}$. As in Theorem 5, we now specify the following parameters which are sufficient to establish the lower bound for GHPM:

1. a sequence (N, k_N) such that k-PDS(N, k_N, p, q) is hard according to Conjecture 3; and
2. a sequence $(n', r', K', \gamma, s, t, \mu)$ with a subsequence that satisfies three conditions: (2.1) the parameters on the subsequence are in the regime of the desired computational lower bound for GHPM; (2.2) the parameters (n', r', K', γ) have the same growth rate as (n, r, K, γ) on this subsequence; and (2.3) such that GHPM(n', r', K', γ) with the parameters on this subsequence can be produced by k-PDS-TO-GHPM with input k-PDS(N, k_N, p, q) applied with additional parameters s, t and μ.

We choose these parameters as follows:

- let $r' = r$ be the smallest prime satisfying that $r \leq r' \leq 2r$, which exists by Bertrand's postulate and can be found in poly(n) time;
- let t be such that $(r')^t$ is the closest power of r' to $r'K/\sqrt{n}$, let $s = \lceil n/r'K \rceil$ and let $\mu = \frac{\gamma (r')^t \sqrt{r'}}{r'-1}$;
- now let k_N be given by

$$k_N = \left\lfloor \frac{1}{2} \left(1 + \frac{p}{Q}\right)^{-1} w^{-2} \cdot \min\left\{ \frac{K}{(r')^{t-1}}, \sqrt{n} \right\} \right\rfloor$$

  where $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}} \left(\sqrt{q} - 1\right)$; and
- let $K' = k_N (r')^{t-1}$, let $n' = k_N s (r')^t$ and let $N = w k_N^2$.

Now observe that we have the following bounds

$$\begin{aligned}
n' &\asymp k_N s (r')^t \asymp \left(w^{-2} \cdot \min\left\{ 1, \frac{(r')^{t-1}\sqrt{n}}{K} \right\} \right) n \\
K' &\asymp k_N (r')^{t-1} = \frac{n'}{r's} \asymp \left(w^{-2} \cdot \min\left\{ 1, \frac{(r')^{t-1}\sqrt{n}}{K} \right\} \right) K \\
m &\leq 2 \left(\frac{p}{Q} + 1 \right) w k_N^2 \leq \left(w^{-1} \cdot \min\left\{ \frac{K}{(r')^{t-1}\sqrt{n}}, 1 \right\} \cdot \frac{K}{(r')^{t-1}\sqrt{n}} \right) k_N s (r'-1) \ell \\
k_N s (r'-1) \ell &\leq \text{poly}(N) \\
(r')^2 (K')^2 &\geq \left(\frac{r'K'}{rK} \right)^2 \cdot \frac{n}{n'} \cdot w' n' \\
\mu = \frac{\gamma (r')^t \sqrt{r'}}{r'-1} &\leq \sqrt{\frac{r'}{r}} \cdot \frac{(r')^{t-1}\sqrt{n}}{K} \cdot \frac{2}{(w')^{1/2}\sqrt{\log n}} \\
\gamma^2 &\lesssim \frac{n'}{w' \cdot r' (K')^2 \log n'} \cdot \frac{r'}{r} \cdot \frac{n'}{n} \cdot \frac{(K')^2}{K^2} \cdot \frac{\log n'}{\log n}
\end{aligned}$$

where m is the smallest multiple of k_N larger $\left(\frac{p}{Q}+1\right)N$ and $\ell = \frac{(r')^t-1}{r'-1}$. Now observe that as long as $r'K/\sqrt{n} = \tilde{\Theta}((r')^t)$ then: (2.1) the last inequality above on γ^2 would imply that (n', r', K', γ) is in the desired hard regime; (2.2) the pairs of parameters (n, n'), (K, K') and (r, r') have the same growth rates since $w = n^{o(1)}$ and either $r' = r$ or $r' = \Theta(r) = \omega(1)$; and (2.3) the third through sixth bounds above imply that taking c large enough yields the conditions needed to apply Corollary 88 to yield the desired reduction. By Lemma 80, there is an infinite subsequence of the input parameters such that $r'K/\sqrt{n} = \tilde{\Theta}((r')^t)$, which concludes the proof of the lower bound for GHPM as in Theorems 4 and 5.

The computational lower bound for BHPM follows from the same argument applied to $\mathcal{A}$ from Corollary 94 with the following modification. The conditions in the theorem statement for BHPM add the initial condition that $rK^{4/3} \geq w'n$. The parameter settings above then imply that $k_N\sqrt{r'} = \tilde{o}((r')^{2t})$ holds on the parameter subsequence with $r'K/\sqrt{n} = \tilde{\Theta}((r')^t)$. The same reasoning above then yields the desired computational lower bound for BHPM and completes the proof of the theorem. ■

M.3. Semirandom Single Community Recovery

In this section, we show that the k-PC and k-PDS conjectures with constant edge density imply the PDS Recovery Conjecture under a semirandom adversary in the regime of constant ambient edge density. The PDS Recovery Conjecture and formulations of semirandom single community recovery here are as they were introduced in Sections B.6 and E.3. Our reduction from k-PDS to SEMI-CR is shown in Figure 14. On a high level, our main observation is that an adversary in SEMI-CR with subgraph size k can simulate the problem of detecting for the presence of a hidden ISBM instance on a subgraph with $O(k)$ in an n-vertex Erdős-Rényi graph. Furthermore, combining the Bernoulli rotations step with $K_{3,t}$ as in k-PDS-TO-ISBM with the partition refinement of k-PDS-TO-GHPM can be shown to map to this detection problem. Furthermore, it faithfully recovers the Kesten-Stigum bound from the PDS Recovery Conjecture as opposed to the slower detection rate. The key proofs in this section resemble similar proofs in the previous two sections. We omit details that are similar for brevity.

Before proceeding with the main proofs of this section, we discuss the relationship between our results and the reduction of Cai et al. (2015a). In Cai et al. (2015a), the authors prove a detection-recovery gap in the context of sub-Gaussian submatrix localization based on the hardness of finding a planted k-clique in a random $n/2$-regular graph. This degree-regular formulation of PC was previously considered in Deshpande and Montanari (2015a) and differs in a number of ways from PC. For example, it is unclear how to generate a sample from the degree-regular variant in polynomial time. We remark that the reduction of Cai et al. (2015a), when instead applied the usual formulation of PC produces a matrix with highly dependent entries. Specifically, the sum of the entries of the output matrix has variance n^2/μ where $\mu \ll 1$ is the mean parameter for the submatrix localization instance whereas an output matrix with independent entries of unit variance would have a sum of entries of variance n^2. Note that, in general, any reduction beginning with PC that also preserves the natural H_0 hypothesis cannot show the existence of a detection-recovery gap, as any lower bounds for localization would also apply to detection.

Formally, the goal of this section is to show that the reduction kPDS-TO-SEMI-CR in Figure 14 maps from k-PC and k-PDS to the following distribution under H_1, for a particular choice of μ_1, μ_2

and μ_3 just below the PDS Recovery Conjecture. We remark that k-PDS-TO-SEMI-CR maps to the specific case where $P_0 = 1/2$. This reduction is extended in Corollary 98 to handle $P_0 \neq 1/2$ with $\min\{P_0, 1 - P_0\} = \Omega(1)$.

Definition 95 (Target SEMI-CR Instance) *Given positive integers $k, k' \leq n$ and $P_0, \mu_1, \mu_2, \mu_3 \in (0, 1)$ satisfying that $\mu_1, \mu_2 \leq P_0 \leq 1 - \mu_3$, let $\mathrm{TSI}(n, k, k', P_0, \mu_1, \mu_2, \mu_3)$ be the distribution over $G \in \mathcal{G}_n$ sampled as follows:*

1. *choose two disjoint subsets $S \subseteq [n]$ and $S' \subseteq [n]$ of sizes $|S| = k$ and $|S'| = k'$, respectively, uniformly at random; and*
2. *include the edge $\{i, j\}$ in $E(G)$ independently with probability p_{ij} where*

$$p_{ij} = \begin{cases} P_0 & \text{if } (i,j) \in S'^2 \\ P_0 - \mu_1 & \text{if } (i,j) \in [n]^2 \backslash (S \cup S')^2 \\ P_0 - \mu_2 & \text{if } (i,j) \in S \times S' \text{ or } (i,j) \in S' \times S \\ P_0 + \mu_3 & \text{if } (i,j) \in S^2 \end{cases}$$

Note that this distribution can be produced by a semirandom adversary in $\mathrm{SEMI\text{-}CR}(n, k, P_0 + \mu_3, P_0)$ under H_1 as follows:

1. samples S' of size k' uniformly at random from all k'-subsets of $[n] \backslash S$ where S is the vertex set of the planted dense subgraph; and
2. if the edge $\{i, j\}$ is in $E(G)$, remove it from G independently with probability q_{ij} where

$$q_{ij} = \begin{cases} 0 & \text{if } (i,j) \in S^2 \cup S'^2 \\ \mu_1/P_0 & \text{if } (i,j) \notin (S \cup S')^2 \\ \mu_2/P_0 & \text{if } (i,j) \in S \times S' \text{ or } (i,j) \in S' \times S \end{cases}$$

Note that $\mathcal{G}(n, P_0')$ can be produced by the adversary under H_0 of $\mathrm{SEMI\text{-}CR}(n, k, P_0 + \mu_1, P_0)$ as long as $P_0' \leq P_0$ by removing all edges independently with probability $1 - P_0'/P_0$. Thus it suffices to map to a testing problem between some $\mathrm{TSI}(n, k, k', P_0, \mu_1, \mu_2, \mu_3)$ and $\mathcal{G}(n, P_0')$.

The next theorem establishes our main Markov transition guarantees for the reduction kPDS-TO-SEMI-CR, which map to such a testing problem when $P_0 = 1/2$.

Theorem 96 (Reduction to SEMI-CR) *Let N be a parameter and fix other parameters as follows:*

- Initial k-BPDS Parameters: *k, N, p, q and E as in Theorem 85.*
- Target SEMI-CR Parameters: *$(n, K, 1/2 + \gamma, 1/2)$ where $n = 3ks \cdot \frac{3^t - 1}{2}$ and $K = (3^t - 1)k$ for some parameters $t = t(N), s = s(N) \in \mathbb{N}$ satisfying that*

$$m \leq 3^t ks \leq n \leq \mathrm{poly}(N)$$

where m and Q are as in Theorem 92. The target level of signal γ is given by $\gamma = \Phi\left(\frac{\mu}{3^t}\right) - 1/2$ and the target TSI densities are

$$\mu_1 = \Phi\left(\frac{\mu}{3^{t+1}}\right) - \frac{1}{2} \quad \text{and} \quad \mu_2 = \mu_3 = \Phi\left(\frac{\mu}{3^t}\right) - \frac{1}{2}$$

where $\mu \in (0,1)$ *satisfies that*

$$\mu \leq \frac{1}{2\sqrt{6\log n + 2\log(p-Q)^{-1}}} \cdot \min\left\{\log\left(\frac{p}{Q}\right), \log\left(\frac{1-Q}{1-p}\right)\right\}$$

Let $\mathcal{A}(G)$ *denote* k-PDS-TO-SEMI-CR *applied to the graph* G *with these parameters. Then* $\mathcal{A}$ *runs in* $\mathrm{poly}(N)$ *time and it follows that*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}_E(N,k,p,q)\right), \text{TSI}(n,K,K/2,1/2,\mu_1,\mu_2,\mu_3)\right) = O\left(\frac{k}{\sqrt{N}} + e^{-\Omega(N^2/km)} + (3^t ks)^{-1}\right)$$
$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}(N,q)\right), \mathcal{G}\left(n,1/2-\mu_1\right)\right) = O\left(e^{-\Omega(N^2/km)} + (3^t ks)^{-1}\right)$$

To prove this theorem, we prove a lemma analyzing the Bernoulli rotations step in Figure 14. The proof of this lemma is similar to those of Lemmas 86 and 93. We omit details that are identical. Recall from Section I.1 the definition of the vector $v_{S,F^s,F^o}(M) \in \mathbb{R}^{ab}$ where F^s and F^o are partitions of $[ab]$ into a equally sized parts and S is a set intersecting each F_i^s in exactly one element. Here we extend this definition to sets S intersecting each F_i^s in at most one element, by setting

$$\left(v_{S,F^s,F^o}(M)\right)_{F_i^o} = \begin{cases} M_{\cdot,S\cap F_i^s} & \text{if } S\cap F_i \neq \emptyset \\ 0 & \text{if } S\cap F_i = \emptyset \end{cases}$$

for each $1 \leq i \leq a$. We now can state the approximate Markov transition guarantees for the Bernoulli rotations step of k-PDS-TO-SEMI-CR in this notation.

Lemma 97 (Bernoulli Rotations for SEMI-CR) *Let* F^s *and* F^o *be a fixed partitions of* $[3^t ks]$ *and* $[n]$ *into* ks *parts of size* 3^t *and* $\frac{1}{2}(3^t-1)$*, respectively, and let* $S \subseteq [3^t ks]$ *where* $|S| = k$ *and* $|S \cap F_i^s| \leq 1$ *for each* $1 \leq i \leq ks$*. Let* $\mathcal{A}_3$ *denote Step 3 of* k-PDS-TO-SEMI-CR *with input* M_{PD2} *and output* M_{R}*. Suppose that* p, Q *and* μ *are as in Theorem 96, then it follows that*

$$d_{TV}\Big(\mathcal{A}_3\left(\mathcal{M}_{[3^t ks]\times[3^t ks]}\left(S\times S, \text{Bern}(p), \text{Bern}(Q)\right)\right), \mathcal{L}\left(\frac{2\mu}{3}\cdot v_{S,F^s,F^o}(K_{3,t}) v_{S,F^s,F^o}(K_{3,t})^\top + \mathcal{N}(0,1)^{\otimes n\times n}\right)\Big) = O\left((3^t ks)^{-1}\right)$$
$$d_{TV}\left(\mathcal{A}_3\left(\text{Bern}(Q)^{\otimes 3^t ks\times 3^t ks}\right), \mathcal{N}(0,1)^{\otimes n\times n}\right) = O\left((3^t ks)^{-1}\right)$$

Proof Let (1) and (2) denote the following two cases:

1. $M_{\text{PD2}} \sim \mathcal{M}_{[3^t ks]\times[3^t ks]}\left(S\times S, \text{Bern}(p), \text{Bern}(Q)\right)$; and
2. $M_{\text{PD2}} \sim \text{Bern}(Q)^{\otimes 3^t ks\times 3^t ks}$.

Now define the matrix M'_{R} with independent entries such that

$$M'_{\text{R}} \sim \begin{cases} \frac{2\mu}{3}\cdot v_{S,F^s,F^o}(K_{3,t}) v_{S,F^s,F^o}(K_{3,t})^\top + \mathcal{N}(0,1)^{\otimes n\times n} & \text{if (1) holds} \\ \mathcal{N}(0,1)^{\otimes n\times n} & \text{if (2) holds} \end{cases}$$

Similarly to Lemma 93, Lemmas 26 and 30 yields that under both (1) and (2), we have that

$$d_{\text{TV}}\left(\left(M_{\text{R}}\right)_{F_i^s,F_j^s}, \left(M'_{\text{R}}\right)_{F_i^s,F_j^s}\right) = O\left(3^{2t}\cdot(3^t ks)^{-3}\right)$$

Algorithm k-PDS-TO-SEMI-CR

Inputs: k-PDS instance $G \in \mathcal{G}_N$ with dense subgraph size k that divides N, and the following parameters

- partition E, edge probabilities $0 < q < p \le 1$, $Q \in (0,1)$ and m as in Figure 12
- refinement parameter s and number of vertices $n = 3ks \cdot \frac{3^t-1}{2}$ for some $t \in \mathbb{N}$ satisfy that $m \le 3^t ks \le n \le \text{poly}(N)$
- mean parameter $\mu \in (0,1)$ as in Figure 12

1. *Symmetrize and Plant Diagonals*: Compute $M_{\text{PD1}} \in \{0,1\}^{m \times m}$ and F as in Step 1 of Figure 12.
2. *Pad and Further Partition*: Form M_{PD2} and F' as in Step 2 of Figure 12 modified so that M_{PD2} is a $3^t ks \times 3^t ks$ matrix and each F'_i has size $3^t s$. Let F^s be the partition of $[3^t ks]$ into ks parts of size 3^t by refining F' by splitting each of its parts into s parts of equal size arbitrarily.
3. *Bernoulli Rotations*: Let F^o be a partition of $[n]$ into ks equally sized parts. Now compute the matrix $M_{\text{R}} \in \mathbb{R}^{n \times n}$ as follows:
   (1) For each $i, j \in [k]$, apply TENSOR-BERN-ROTATIONS to the matrix $(M_{\text{P}})_{F_i^s, F_j^s}$ with matrix parameter $A_1 = A_2 = K_{3,t}$, Bernoulli probabilities $0 < Q < p \le 1$, output dimension $\frac{1}{2}(3^t - 1)$, $\lambda_1 = \lambda_2 = \sqrt{3/2}$ and mean parameter μ.
   (2) Set the entries of $(M_{\text{R}})_{F_i^o, F_j^o}$ to be the entries in order of the matrix output in (1).
4. *Threshold and Output*: Output the graph generated by Step 4 of Figure 12 modified so that G' has vertex set $[n]$ and M_{R} is thresholded at $\frac{\mu}{3^{t+1}}$.

Figure 14: Reduction from k-partite planted dense subgraph to semirandom community recovery.

for all $1 \le i, j \le ks$. The tensorization property of total variation in Fact 15 now yields that

$$d_{\text{TV}}\left(\mathcal{L}(M_{\text{R}}), \mathcal{L}(M'_{\text{R}})\right) = O\left((3^t ks)^{-1}\right)$$

under both (1) and (2), proving the lemma. ■

We now complete the proof of Theorem 96, which follows a similar structure as in Theorem 85.

Proof [Proof of Theorem 96] Let the steps of $\mathcal{A}$ to map inputs to outputs as follows

$$(G, E) \xrightarrow{\mathcal{A}_1} (M_{\text{PD1}}, F) \xrightarrow{\mathcal{A}_2} (M_{\text{PD2}}, F^s) \xrightarrow{\mathcal{A}_3} (M_{\text{R}}, F^o) \xrightarrow{\mathcal{A}_4} G'$$

Under H_1, consider Lemma 16 applied to the following sequence of distributions

$$\mathcal{P}_0 = \mathcal{G}_E(N, k, p, q)$$

$$\mathcal{P}_1 = \mathcal{M}_{[m]\times[m]}(S \times S, \text{Bern}(p), \text{Bern}(Q)) \quad \text{where } S \sim \mathcal{U}_m(F)$$
$$\mathcal{P}_2 = \mathcal{M}_{[3^t ks]\times[3^t ks]}(S \times S, \text{Bern}(p), \text{Bern}(Q)) \quad \text{where } S \sim \mathcal{U}^k_{3^t ks}(F^s)$$
$$\mathcal{P}_3 = \frac{2\mu}{3} \cdot v_{S,F^s,F^o}(K_{3,t}) v_{S,F^s,F^o}(K_{3,t})^\top + \mathcal{N}(0,1)^{\otimes n\times n} \quad \text{where } S \sim \mathcal{U}^k_{3^t ks}(F^s)$$
$$\mathcal{P}_4 = \text{TSI}(n, K, K/2, 1/2, \mu_1, \mu_2, \mu_3)$$

Let $C_Q = \max\left\{\frac{Q}{1-Q}, \frac{1-Q}{Q}\right\}$ and consider setting

$$\epsilon_1 = 4k \cdot \exp\left(-\frac{Q^2N^2}{48pkm}\right) + \sqrt{\frac{C_Q k^2}{2m}}, \quad \epsilon_2 = 0, \quad \epsilon_3 = O\left((3^t ks)^{-1}\right) \quad \text{and} \quad \epsilon_4 = 0$$

Lemma 23 implies this is a valid choice of ϵ_1, $\mathcal{A}_2$ is exact so we can take $\epsilon_2 = 0$ and ϵ_3 is valid by applying Lemma 97 and averaging over $S \sim \mathcal{U}^k_{3^t ks}(F^s)$ using the conditioning property of total variation in Fact 15. Now note that for each S the definition of $v_{S,F^s,F^o}(K_{3,t})$ implies that there are sets S_1 and S_2 with $|S_1| = (3^t - 1)k$ and $|S_2| = \frac{3^t-1}{2} \cdot k$ such that

$$\left(\frac{2\mu}{3} \cdot v_{S,F^s,F^o}(K_{3,t}) v_{S,F^s,F^o}(K_{3,t})^\top\right)_{ij} = \frac{\mu}{3^{t+1}} + \begin{cases} \mu/3^t & \text{if } i,j \in S_1 \\ -\mu/3^t & \text{if } (i,j) \in S_1 \times S_2 \text{ or } (i,j) \in S_2 \times S_1 \\ 0 & \text{if } i,j \in S_2 \\ -\mu/3^{t+1} & \text{if } i,j \notin (S_1 \cup S_2) \end{cases}$$

for each $1 \le i, j \le n$. Permuting the rows and columns of $\mathcal{P}_3$ therefore yields $\mathcal{P}_4$ exactly with $\epsilon_4 = 0$. Lemma 16 thus establishes the first bound. Under H_0, consider the distributions

$$\mathcal{P}_0 = \mathcal{G}(N, q), \quad \mathcal{P}_1 = \text{Bern}(Q)^{\otimes m\times m}, \quad \mathcal{P}_2 = \text{Bern}(Q)^{\otimes 3^t ks\times 3^t ks},$$
$$\mathcal{P}_3 = \mathcal{N}(0,1)^{\otimes n\times n} \quad \text{and} \quad \mathcal{P}_4 = \mathcal{G}\left(n, 1/2 - \mu_1\right)$$

As in Theorems 85 and 92, Lemmas 23 and 97 imply $\epsilon_1 = 4k \cdot \exp\left(-\frac{Q^2N^2}{48pkm}\right)$ and the choices of ϵ_2, ϵ_3 and ϵ_4 above are valid. Lemma 16 now yields the second bound and completes the proof of the theorem. ■

We now add a simple final step to kPDS-TO-SEMI-CR, reducing to arbitrary $P_0 \neq 1/2$. The guarantees for this modified reduction are captured in the following corollary.

Corollary 98 (Arbitrary Bounded P_0) *Define all parameters as in Theorem 96 and let $P_0 \in (0,1)$ be such that $\eta = \min\{P_0, 1 - P_0\} = \Omega(1)$. Then there is a* poly$(N)$ *time reduction $\mathcal{A}$ from graphs on N vertices to graphs on n vertices satisfying that*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}_E(N,k,p,q)\right), \text{TSI}(n, K, K/2, P_0, 2\eta\mu_1, 2\eta\mu_2, 2\eta\mu_3)\right) = O\left(\frac{k}{\sqrt{N}} + e^{-\Omega(N^2/km)} + (3^t ks)^{-1}\right)$$
$$d_{TV}\left(\mathcal{A}\left(\mathcal{G}(N,q)\right), \mathcal{G}\left(n, P_0 - 2\eta\mu_1\right)\right) = O\left(e^{-\Omega(N^2/km)} + (3^t ks)^{-1}\right)$$

Proof This corollary follows from the same reduction in the first part of the proof of Corollary 88. Consider the reduction $\mathcal{A}$ that adds a simple post-processing step to k-PDS-TO-SEMI-CR as follows. On input graph G with N vertices:

1. Form the graph G_1 by applying k-PDS-TO-CR to G with parameters N, k, E, ℓ, n, s, t and μ.
2. Form G_2 as in $\mathcal{A}_2$ of Corollary 88.

This clearly runs in poly(N) time and the second step can be verified to map $\text{TSI}(n, K, K/2, 1/2, \mu_1, \mu_2, \mu_3)$ to $\text{TSI}(n, K, K/2, P_0, 2\eta\mu_1, 2\eta\mu_2, 2\eta\mu_3)$ and $\mathcal{G}(n, 1/2 - \mu_1)$ to $\mathcal{G}(n, P_0 - 2\eta\mu_1)$ exactly. Applying Theorem 96 and Lemma 16 to each of these two steps proves the bounds in the corollary statement. ■

Summarizing the results of this section, we arrive at the desired computational lower bound for SEMI-CR. The proof of the next theorem follows the usual recipe for deducing computational lower bounds and is deferred to Appendix R.2.

Theorem 12 (Lower Bounds for SEMI-CR) *If k and n are polynomial in each other with $k = \Omega(\sqrt{n})$ and $0 < P_0 < P_1 \le 1$ where $\min\{P_0, 1 - P_0\} = \Omega(1)$, then the k-PC conjecture or k-PDS conjecture for constant $0 < q < p \le 1$ both imply that there is a computational lower bound for $\text{SEMI-CR}(n, k, P_1, P_0)$ at $\frac{(P_1 - P_0)^2}{P_0(1-P_0)} = \tilde{o}(n/k^2)$.*

Appendix N. Tensor Principal Component Analysis

In this section, we: (1) give our reduction k-PST-TO-TPCA from k-partite planted sub-tensor to tensor PCA; (2) combine this with the completing hypergraphs technique of Section J to prove our main computational lower bound for the hypothesis testing formulation of tensor PCA, Theorem 10; and (3) we show that Theorem 10 implies computational lower bounds for the recovery formulation of tensor PCA. We remark that the heuristic at the end of Section C.3 yields the predicted computational barrier for TPCA. Specifically, the ℓ_2 norm for the data tensor $\mathbb{E}[X]$ corresponding to k-HPC^s is $\Theta(k^{s/2})$ which is $\tilde{\Theta}(n^{s/4})$ just below the conjectured computational barrier for k-HPC^s. Furthermore, the corresponding ℓ_2 norm for H_1 of TPCA^s is $\tilde{\Theta}(\theta n^{s/2})$. Equating these norms correctly predicts the computational barrier of $\theta = \tilde{\Theta}(n^{-s/4})$.

Our reduction k-PST-TO-TPCA is shown in Figure 15, which applies dense Bernoulli rotations with Kronecker products of the matrices $K_{2,t}$ to the planted sub-tensor problem. The following theorem establishes the approximate Markov transition properties of this reduction. Its proof is similar to the proofs of Theorems 46 and 85. We omit details that are similar for brevity.

Theorem 99 (Reduction to Tensor PCA) *Fix initial and target parameters as follows:*

- Initial k-PST Parameters: *dimension N, sub-tensor size k that divides N, order s, a partition F of $[N]$ into k parts of size N/k and edge probabilities $0 < q < p \le 1$ where $\min\{q, 1 - q\} = \Omega_N(1)$.*
- Target TPCA Parameters: *dimension n and a parameter $t = t(N) \in \mathbb{N}$ satisfying that*

$$n \le D = 2k(2^t - 1), \quad N \le 2^t k \quad \text{and} \quad t = O(\log N)$$

and target level of signal $\theta \in (0, 1)$ where

$$\theta \le \frac{c \cdot \delta}{2^{st/2} \cdot \sqrt{t + \log(p - q)^{-1}}}$$

for a sufficiently small constant $c > 0$, where $\delta = \min\left\{\log\left(\frac{p}{q}\right), \log\left(\frac{1-q}{1-p}\right)\right\}$.

Algorithm k-PST-TO-TPCA

Inputs: k-PST instance $T \in \{0,1\}^{N^{\otimes s}}$ of order s with planted sub-tensor size k that divides N, and the following parameters

- partition F of $[N]$ into k parts of size N/k and edge probabilities $0 < q < p \leq 1$
- output dimension n and a parameter $t \in \mathbb{N}$ satisfying that

$$n \leq D = 2k(2^t - 1), \quad N \leq 2^t k \quad \text{and} \quad t = O(\log N)$$

- target level of signal $\theta \in (0,1)$ where

$$\theta \leq \frac{c \cdot \delta}{2^{st/2} \cdot \sqrt{t + \log(p-q)^{-1}}}$$

for a sufficiently small constant $c > 0$, where $\delta = \min\left\{\log\left(\frac{p}{q}\right), \log\left(\frac{1-q}{1-p}\right)\right\}$.

1. *Pad*: Form $T_{\text{PD}} \in \{0,1\}^{2^t k \times 2^t k}$ by embedding T as the upper left principal sub-tensor of T_{PD} and then adding $2^t k - N$ new indices along each axis of T and filling all missing entries with i.i.d. samples from $\text{Bern}(q)$. Let F_i' be F_i with $2^t - N/k$ of the new indices. Sample k random permutations σ_i of F_i' independently for each $1 \leq i \leq k$ and permute the indices along each axis of T_{PD} within each part F_i' according to σ_i.
2. *Bernoulli Rotations*: Let F'' be a partition of $[D]$ into k equally sized parts. Now compute the matrix $T_{\text{R}} \in \mathbb{R}^{K^{\otimes s}}$ as follows:
   (1) For each block index $(i_1, i_2, \ldots, i_s) \in [k]$, apply TENSOR-BERN-ROTATIONS to the tensor $(T_{\text{PD}})_{F_{i_1}', F_{i_2}', \ldots, F_{i_s}'}$ with matrix parameters $A_1 = A_2 = \cdots = A_s = K_{2,t}$, rejection kernel parameter $R_{\text{RK}} = (2^t k)^s$, Bernoulli probabilities $0 < Q < p \leq 1$, output dimension $D/k = 2(2^t - 1)$, singular value upper bounds $\lambda_1 = \lambda_2 = \cdots = \lambda_s = \sqrt{2}$ and mean parameter $\mu = \theta \cdot 2^{s(t+1)/2}$.
   (2) Set the entries of $(T_{\text{R}})_{F_{i_1}'', F_{i_2}'', \ldots, F_{i_s}''}$ to be the entries in order of the tensor output in (1).
3. *Subsample, Sign and Output*: Randomly choose a subset $U \subseteq [D]$ of size $|U| = n$ and randomly sample a vector $b \sim \text{Unif}\,[\{-1,1\}]^{\otimes D}$ output the tensor $b^{\otimes s} \odot T_{\text{R}}$ restricted to the indices in U, or in other words $(b^{\otimes s} \odot T_{\text{R}})_{U,U,\ldots,U}$, where $\odot$ denotes the entrywise product of two tensors.

Figure 15: Reduction from k-partite Bernoulli planted sub-tensor to tensor PCA.

Let $\mathcal{A}(T)$ denote k-PST-TO-TPCA applied to the tensor T with these parameters. Then $\mathcal{A}$ runs in $\text{poly}(N)$ *time and it follows that*

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[N]^s}\left(S^s, \text{Bern}(p), \text{Bern}(q)\right)\right), \text{TPCA}_D^s(n,\theta)\right) = O\left(k^{-2s}2^{-2st}\right)$$

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[N]^s}\left(\text{Bern}(q)\right)\right), \mathcal{N}(0,1)^{\otimes n^{\otimes s}}\right) = O\left(k^{-2s}2^{-2st}\right)$$

for any set $S \subseteq [N]$ *with* $|S \cap E_i| = 1$ *for each* $1 \le i \le k$.

We now prove two lemmas stating the guarantees for the dense Bernoulli rotations step and final step of k-PST-TO-TPCA. Define $v_{S,F',F''}(M)$ as in Section I.1. Note that the matrix $K_{2,t}$ has dimensions $2(2^t - 1) \times 2^t$. The proof of the next lemma follows from the same argument as in the proof of Lemma 47.

Lemma 100 (Bernoulli Rotations for TPCA) *Let* F' *and* F'' *be fixed partitions of* $[2^t k]$ *and* $[D]$ *into* k *parts of size* 2^t *and* $2(2^t - 1)$*, respectively, and let* $S \subseteq [2^t k]$ *where* $|S \cap F'_i| = 1$ *for each* $1 \le i \le k$. *Let* $\mathcal{A}_2$ *denote Step 2 of* k-PST-TO-TPCA *with input* T_{PD} *and output* T_{R}. *Suppose that* p, q *and* θ *are as in Theorem 99, then it follows that*

$$d_{TV}\Big(\mathcal{A}_2\left(\mathcal{M}_{[2^t k]^s}\left(S^s, \mathrm{Bern}(p), \mathrm{Bern}(q)\right)\right), \\ \mathcal{L}\left(2^{st/2}\theta \cdot v_{S,F',F''}(K_{2,t})^{\otimes s} + \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right)\Big) = O\left(k^{-2s}2^{-2st}\right)$$

$$d_{TV}\left(\mathcal{A}_2\left(\mathcal{M}_{[2^t k]^s}\left(\mathrm{Bern}(q)\right)\right), \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right) = O\left(k^{-2s}2^{-2st}\right)$$

Proof This lemma follows from the same argument as in the proof of Lemma 47. We outline the details that differ. Specifically, consider the case in which $T_{\mathrm{PD}} \sim \mathcal{M}_{[2^t k]^s}\left(S^s, \mathrm{Bern}(p), \mathrm{Bern}(q)\right)$. Observe that

$$(T_{\mathrm{PD2}})_{F'_{i_1}, F'_{i_2}, \ldots, F'_{i_s}} \sim \mathrm{PB}\left(F'_{i_1} \times F'_{i_2} \times \cdots \times F'_{i_s}, (S \cap F'_{i_1}, S \cap F'_{i_2}, \ldots, S \cap F'_{i_s}), p, q\right)$$

for all $(i_1, i_2, \ldots, i_s) \in [k]^s$. The singular value upper bound on $K_{2,t}$ in Lemma 30 and the same application of Corollary 27 as in Lemma 47 yields that

$$d_{\mathrm{TV}}\left((T_{\mathrm{R}})_{F''_{i_1}, \ldots, F''_{i_s}}, \mathcal{L}\left(2^{-s/2}\mu \cdot (K_{2,t})_{\cdot, S\cap F'_{i_1}} \otimes \cdots \otimes (K_{2,t})_{\cdot, S \cap F'_{i_s}} + \mathcal{N}(0,1)^{\otimes (D/k)^{\otimes s}}\right)\right) \\ = O\left(k^{-3s}2^{-2st}\right)$$

for all $(i_1, i_2, \ldots, i_s) \in [k]^s$ since $\prod_{j=1}^s \lambda_j = 2^{s/2}$. Note that the exponent of 8 is guaranteed by changing the parameter in Gaussian rejection kernels from n to n^{10} to decrease their total variation error. Note that this step still runs in $\mathrm{poly}(n^{10})$ time. Combining this bound for all such $(i_1, i_2, \ldots, i_s)$ and the tensorization property of total variation in Fact 15 yields that

$$d_{\mathrm{TV}}\left(T_{\mathrm{R}}, \mathcal{L}\left(2^{-s/2}\mu \cdot v_{S,F',F''}(K_{2,t})^{\otimes s} + \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right)\right) = O\left(k^{-2s}2^{-2st}\right)$$

Combining this with the fact that $\mu = \theta \cdot 2^{s(t+1)/2}$ now yields the first bound in the lemma. The second bound follows by the same argument but now applying Corollary 27 to the distribution $(T_{\mathrm{PD2}})_{F'_{i_1}, \ldots, F'_{i_s}} \sim \mathrm{Bern}(q)^{(D/k)^{\otimes s}}$. This completes the proof of the lemma. ■

Lemma 101 (Signing for TPCA) *Let* F', F'' *and* S *be as in Lemma 100 and let* p, q *and* θ *be as in Theorem 99. Let* $\mathcal{A}_3$ *denote Step 3 of* k-PST-TO-TPCA *with input* T_{R} *and output given by the output* T' *of* $\mathcal{A}$. *Then*

$$\mathcal{A}_3\left(2^{st/2}\theta \cdot v_{S,F',F''}(K_{2,t})^{\otimes s} + \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right) \sim \mathrm{TPCA}^s_D(n, \theta)$$

$$\mathcal{A}_3\left(\mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right) \sim \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$$

Proof Suppose that $T_{\mathrm{R}} \sim \mathcal{L}\left(2^{st/2}\theta \cdot v_{S,F',F''}(K_{2,t})^{\otimes s} + \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right)$ and let $b \sim \mathrm{Unif}\left[\{-1,1\}\right]^{\otimes D}$ be as in Step 3 of $\mathcal{A}$. The symmetry of zero-mean Gaussians and independence among the entries of $\mathcal{N}(0,1)^{\otimes D^{\otimes s}}$ imply that

$$b^{\otimes s} \odot T_{\mathrm{R}} \sim \mathcal{L}\left(2^{st/2}\theta \cdot u^{\otimes s} + b^{\otimes s} \odot \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right) = \mathcal{L}\left(2^{st/2}\theta \cdot u^{\otimes s} + \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right)$$

where $u = b \odot v_{S,F',F''}(K_{2,t})$ and the two terms $u^{\otimes s}$ and $\mathcal{N}(0,1)^{\otimes D^{\otimes s}}$ above are independent. Now note that each entry of $v_{S,F',F''}(K_{2,t})$ is either $\pm 2^{-t/2}$ by the definition of $K_{2,t}$. This implies that $2^{t/2}u$ is distributed as $\mathrm{Unif}\left[\{-1,1\}\right]^{\otimes D}$ and hence that

$$\mathcal{L}\left(b^{\otimes s} \odot T_{\mathrm{R}}\right) = \mathcal{L}\left(\theta \cdot b^{\otimes s} + \mathcal{N}(0,1)^{\otimes D^{\otimes s}}\right) = \mathrm{TPCA}^s_D(D,\theta)$$

Subsampling the same set U of n coordinates of this tensor along each axis by definition yields $\mathrm{TPCA}(n,\theta)$, proving the first claim in the lemma. The second claim is immediate by the fact that if $T_{\mathrm{R}} \sim \mathcal{N}(0,1)^{\otimes D^{\otimes s}}$ then it also holds that $b^{\otimes s} \odot T_{\mathrm{R}} \sim \mathcal{N}(0,1)^{\otimes D^{\otimes s}}$. This completes the proof of the lemma. ■

We now complete the proof of Theorem 99 by applying Lemma 16 as in Theorems 46 and 85.

Proof [Proof of Theorem 99] Define the steps of $\mathcal{A}$ to map inputs to outputs as follows

$$(T,F) \xrightarrow{\mathcal{A}_1} (T_{\mathrm{PD}},F) \xrightarrow{\mathcal{A}_2} (T_{\mathrm{R}},F'') \xrightarrow{\mathcal{A}_3} T'$$

Consider Lemma 16 applied to the following sequence of distributions

$$\begin{aligned}
\mathcal{P}_0 &= \mathcal{M}_{[N]^s}\left(S^s, \mathrm{Bern}(p), \mathrm{Bern}(q)\right) \\
\mathcal{P}_1 &= \mathcal{M}_{[2^t k]^s}\left(S^s, \mathrm{Bern}(p), \mathrm{Bern}(q)\right) \quad \text{where } S \sim \mathcal{U}_{2^t k}(F') \\
\mathcal{P}_2 &= 2^{st/2}\theta \cdot v_{S,F',F''}(K_{2,t})^{\otimes s} + \mathcal{N}(0,1)^{\otimes D^{\otimes s}} \quad \text{where } S \sim \mathcal{U}_{2^t k}(F') \\
\mathcal{P}_3 &= \mathrm{TPCA}^s_D(n,\theta)
\end{aligned}$$

Consider applying Lemmas 100 and 101 while averaging over $S \sim \mathcal{U}_{2^t k}(F')$ and applying the conditioning property of total variation in Fact 15. This yields that we may take $\epsilon_1 = 0$, $\epsilon_2 = O\left(k^{-2s}2^{-2st}\right)$ and $\epsilon_3 = 0$. Applying Lemma 16 proves the first bound in the theorem. Now consider the following sequence of distributions

$$\mathcal{P}_0 = \mathcal{M}_{[N]^s}\left(\mathrm{Bern}(q)\right), \quad \mathcal{P}_1 = \mathcal{M}_{[2^t k]^s}\left(\mathrm{Bern}(q)\right), \quad \mathcal{P}_2 = \mathcal{N}(0,1)^{\otimes D^{\otimes s}} \quad \text{and} \quad \mathcal{P}_3 = \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$$

Lemmas 100 and 101 imply we can again take $\epsilon_1 = 0$, $\epsilon_2 = O\left(k^{-2s}2^{-2st}\right)$ and $\epsilon_3 = 0$. The second bound in the theorem now follows from Lemma 16. ■

We now apply this theorem to deduce our main computational lower bounds for tensor PCA by verifying its guarantees are sufficient to apply Lemma 60.

Theorem 10 (Lower Bounds for TPCA) *Let n be a parameter and $s \geq 3$ be a constant, then the k-HPC^s or k-HPDS^s conjecture for constant $0 < q < p \leq 1$ both imply a computational lower bound for $\mathrm{TPCA}^s(n,\theta)$ at all levels of signal $\theta = \tilde{o}(n^{-s/4})$ against $\mathrm{poly}(n)$ time algorithms $\mathcal{A}$ solving $\mathrm{TPCA}^s(n,\theta)$ with a low false positive probability of $\mathbb{P}_{H_0}[\mathcal{A}(T) = H_1] = O(n^{-s})$.*

Proof We will verify that the approximate Markov transition guarantees for k-PST-TO-TPCA in Theorem 99 are sufficient to apply Lemma 60 for the set of $\mathcal{P} = \text{TPCA}^s(n, \theta)$ with parameters (n, θ) that fill out the region $\theta = \tilde{o}(n^{-s/4})$. Fix a constant pair of probabilities $0 < Q < p \leq 1$, a constant positive integer s and any sequence of parameters (n, θ) where $\theta \in (0, 1)$ is implicitly a function of n with

$$\theta \leq \frac{c}{w^{s/2} n^{s/4} \sqrt{\log n}}$$

for sufficiently large n, an arbitrarily slow-growing function $w = w(n) \to \infty$ and a sufficiently small constant $c > 0$. Now consider the parameters (N, k) and input t to k-PST-TO-TPCA defined as follows:

- let t be such that 2^t is the smallest power of two greater than $w\sqrt{n}$; and
- let $k = \lceil w^{-1}\sqrt{n} \rceil$ and let N be the largest multiple of k less than n.

Now observe that these choices of parameters ensure that k divides N, it holds that $k = o(\sqrt{N})$ and

$$N \leq n \leq 2^t k \leq D = 2k(2^t - 1)$$

Furthermore, we have that $N = \Theta(n)$ and $2^t = \Theta(w\sqrt{n})$. For a sufficiently small choice of $c > 0$, we also have that

$$\theta \leq \frac{c}{w^{s/2} n^{s/4} \sqrt{\log n}} \leq \frac{c' \cdot \delta}{2^{st/2} \cdot \sqrt{t + \log(p - Q)^{-1}}}$$

where $c' > 0$ is the constant and δ is as in Theorem 99. This verifies all of the conditions needed to apply Theorem 99, which implies that k-PST-TO-TPCA maps $k\text{-PST}^s_E(N, k, p, Q)$ to $\text{TPCA}^s(n, \theta)$ under both H_0 and H_1 to within total variation error $O\left(k^{-2s} 2^{-2st}\right) = O(n^{-2s})$. By Lemma 60, the k-HPDS^s conjecture for $k\text{-HPDS}^s_E(N', k', p, q)$ where $N = N' - (s-1)N'/k'$ and $k = k' - s + 1$ now implies that there are is no $\text{poly}(n)$ time algorithm $\mathcal{A}$ solving $\text{TPCA}^s(n, \theta)$ with a low false positive probability of $\mathbb{P}_{H_0}[\mathcal{A}(T) = H_1] = O(n^{-s})$. This completes the proof of the theorem. ■

We conclude this section with the following lemma observing that this theorem implies a computational lower bound for estimating v in $\text{TPCA}^s(n, \theta)$ where $\theta = \tilde{\omega}(n^{-s/2})$ and $\theta = \tilde{o}(n^{-s/4})$. Note that the requirement $\theta = \tilde{\omega}(n^{-s/2})$ is weaker than the condition $\theta = \tilde{\omega}(n^{(1-s)/2})$, which is necessary for recovering v to be information-theoretically possible, as discussed in Section B.10. The next lemma shows that any estimator yields a test in the hypothesis testing formulation of tensor PCA that must have a low false positive probability of error, since thresholding $\langle \hat{v}, T \rangle$ where $\hat{v}$ is an estimator of v, yields a means to distinguish H_0 and H_1 with high probability. We remark that the requirement $\langle v, \hat{v} \rangle = \Omega(\|v\|_2)$ is weaker than the condition $\|v - \hat{v} \cdot \sqrt{n}\|_2 = o(\sqrt{n})$ when $\hat{v}$ is a unit vector and $v \in \{-1, 1\}^n$. Thus any estimation algorithm with ℓ_2 error $o(\sqrt{n})$, directly yields an algorithm $\mathcal{A}_E$ satisfying the conditions of the lemma.

Lemma 102 (One-Sided Blackboxes from Estimation in Tensor PCA) *Let $s \geq 2$ be a fixed constant and suppose that there is a $\text{poly}(n)$ time algorithm $\mathcal{A}_E$ that, on input sampled from $\theta v^{\otimes s} + \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$ where $v \in \{-1, 1\}^n$ is fixed but unknown to $\mathcal{A}_E$ and $\theta = \omega(n^{-s/2}\sqrt{s \log n})$, outputs a unit vector $\hat{v} \in \mathbb{R}^n$ with $\langle v, \hat{v} \rangle = \Omega(\|v\|_2)$. Then there is a $\text{poly}(n)$ time algorithm $\mathcal{A}_D$ solving $\text{TPCA}^s(n, \theta)$ with a low false positive probability of $\mathbb{P}_{H_0}[\mathcal{A}_D(T) = H_1] = O(n^{-s})$.*

Proof Let T be an instance of $\text{TPCA}^s(n,\theta)$ with $T = \theta v^{\otimes s} + G$ under H_1 and $T = G$ under H_0 where $G \sim \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$. Consider the following algorithm $\mathcal{A}_D$ for $\text{TPCA}^s(n,\theta)$:

1. Independently sample $G' \sim \mathcal{N}(0,1)^{\otimes n^{\otimes s}}$ and form $T_1 = \frac{1}{\sqrt{2}}(T + G')$ and $T_2 = \frac{1}{\sqrt{2}}(T - G')$.
2. Compute $\hat{v}(T_1)$ as the output of $\mathcal{A}_E$ applied to T_1.
3. Output H_0 if $\langle \hat{v}(T_1)^{\otimes s}, T_2 \rangle < 2\sqrt{s \log n}$ and output H_1 otherwise.

First note that the entries of $\frac{1}{\sqrt{2}}(G + G')$ and $\frac{1}{\sqrt{2}}(G - G')$ are jointly Gaussian but uncorrelated, which implies that these two tensors are independent. This implies that T_1 and T_2 are independent. Since $\hat{v}(T_1)$ is a unit vector and independent of T_2, it follows that $\langle \hat{v}(T_1)^{\otimes s}, T_2 \rangle$ is distributed as $\mathcal{N}(0,1)$ conditioned on $\hat{v}(T_1)$ if T is distributed according to H_0 of $\text{TPCA}^s(n,\theta)$. Now we have that

$$\mathbb{P}_{H_0}[\mathcal{A}_D(T) = H_1] = \mathcal{P}\left[\langle \hat{v}(T_1)^{\otimes s}, T_2 \rangle \geq 2\sqrt{s \log n}\right] = O(n^{-2s})$$

where the second equality follows from standard Gaussian tail bounds. If T is distributed according to H_1, then $\langle \hat{v}(T_1)^{\otimes s}, T_2 \rangle \sim \mathcal{N}(\theta \langle \hat{v}(T_1), v \rangle^s, 1)$. In this case, $\mathcal{A}_E$ ensures that $\langle \hat{v}(T_1), v \rangle^s = \Omega(n^{s/2})$ since $\|v\|_2 = \sqrt{n}$, and therefore $\theta \langle \hat{v}(T_1), v \rangle^s = \omega(\sqrt{s \log n})$. It therefore follows that

$$\mathbb{P}_{H_1}[\mathcal{A}_D(T) = H_0] \leq \mathcal{P}\left[\langle \hat{v}(T_1)^{\otimes s}, T_2 \rangle - \theta \langle \hat{v}(T_1), v \rangle^s < -2\sqrt{s \log n}\right] = O(n^{-2s})$$

Thus $\mathcal{A}_D$ has Type I+II error that is $o(1)$ and the desired low false positive probability, which completes the proof of the lemma. ■

Appendix O. Universality of Lower Bounds for Learning Sparse Mixtures

In this section, we combine our reduction to ISGM from Section I.1 with symmetric 3-ary rejection kernels, which were introduced and analyzed in Section F.3. We remark that the k-partite promise in k-PDS is crucially used in our reduction to obtain this universality. In particular, this promise ensures that the entries of the intermediate ISGM instance are from one of three distinct distributions, when conditioned on the part of the mixture the sample is from. This is necessary for our application of symmetric 3-ary rejection kernels. An overview of the ideas in this section can be found in Section C.7.

Our general lower bound holds given tail bounds on the likelihood ratios between the planted and noise distributions, and applies to a wide range of natural distributional formulations of learning sparse mixtures. For example, our general lower bound recovers the tight computational lower bounds for sparse PCA in the spiked covariance model from Gao et al. (2017), Brennan et al. (2018) and Brennan and Bresler (2019). The results in this section can also be interpreted as a universality principle for computational lower bounds in sparse PCA. We prove the approximate Markov transition guarantees for our reduction to GLSM in Section O.1 and discuss the universality conditions needed for our lower bounds in Section O.2.

Algorithm k-BPDS-TO-GLSM

Inputs: Matrix $M \in \{0,1\}^{m \times n}$, dense subgraph dimensions k_m and k_n where k_n divides n and the following parameters

- partition F, edge probabilities $0 < q < p \le 1$ and $w(n)$ as in Figure 9
- target GLSM parameters (N, k_m, d) satisfying $wN \le n$ and $m \le d$, a mixture distribution $\mathcal{D}$ and target distributions $\{\mathcal{P}_\nu\}_{\nu \in \mathbb{R}}$ and $\mathcal{Q}$

1. *Map to Gaussian Sparse Mixtures*: Form the sample $Z_1, Z_2, \dots, Z_N \in \mathbb{R}^d$ by setting

$$(Z_1, Z_2, \dots, Z_N) \leftarrow k\text{-BPDS-TO-ISGM}(M, F)$$

   where k-BPDS-TO-ISGM is applied with $r = 2$, slow-growing function $w(n)$, $t = \lceil \log_2(n/k_n) \rceil$, target parameters (N, k_m, d), $\epsilon = 1/2$ and $\mu = c_1 \sqrt{\frac{k_n}{n \log n}}$ for a sufficiently small constant $c_1 > 0$.

2. *Truncate and 3-ary Rejection Kernels*: Sample $\nu_1, \nu_2, \dots, \nu_N \sim_{\text{i.i.d.}} \mathcal{D}$, truncate the ν_i to lie within $[-1, 1]$ and form the vectors $X_1, X_2, \dots, X_N \in \mathbb{R}^d$ by setting

$$X_{ij} \leftarrow 3\text{-SRK}(\text{TR}_\tau(Z_{ij}), \mathcal{P}_{\nu_i}, \mathcal{P}_{-\nu_i}, \mathcal{Q})$$

   for each $i \in [N]$ and $j \in [d]$. Here 3-SRK is applied with $N_{\text{it}} = \lceil 4 \log(dN) \rceil$ iterations and with the parameters

$$a = \Phi(\tau) - \Phi(-\tau), \quad \mu_1 = \frac{1}{2}\left(\Phi(\tau + \mu) - \Phi(\tau - \mu)\right),$$
$$\mu_2 = \frac{1}{2}\left(2 \cdot \Phi(\tau) - \Phi(\tau + \mu) - \Phi(\tau - \mu)\right)$$

3. *Output*: The vectors $(X_1, X_2, \dots, X_N)$.

Figure 16: Reduction from k-part bipartite planted dense subgraph to general learning sparse mixtures.

O.1. Reduction to Generalized Learning Sparse Mixtures

In this section, we combine symmetric 3-ary rejection kernels with the reduction k-BPDS-TO-ISGM to map from k-BPDS to generalized sparse mixtures. The details of this reduction k-BPDS-TO-GLSM are shown in Figure 16. As mentioned in Sections C.7 and F.3, to reduce to sparse mixtures near their computational barrier, it is crucial to produce multiple planted distributions. Previous rejection kernels do not have enough degrees of freedom to map to three output distributions given their binary inputs. The symmetric 3-ary rejection kernels introduced in Section F.3 overcome this issue by mapping three input to three output distributions. In particular, we will see in this section that their approximate Markov transition guarantees established in Lemma 25 exactly lead to tight

computational lower bounds for GLSM. Throughout this section, we will adopt the definitions of GLSM and GLSM_D introduced in Sections B.11 and E.3.

In order to establish computational lower bounds for GLSM, it is crucial to define a meaningful notion of the level of signal in a set of target distributions $\mathcal{D}, \mathcal{Q}$ and $\{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}$. This level of signal was defined in Section B.11 and is reviewed below for convenience. We remark that this definition will turn out to coincide with the conditions needed to apply symmetric 3-ary rejection kernels. This notion of signal also implicitly defines the universality class over which our computational lower bounds hold.

Definition 13 (Universal Class and Level of Signal) *Given a parameter N, define the collection of distributions $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})$ implicitly parameterized by N to be in the universality class $\text{UC}(N)$ if*

- *the pairs $(\mathcal{P}_\nu, \mathcal{Q})$ are all computable pairs, as in Definition 24, for all $\nu \in \mathbb{R}$;*
- *$\mathcal{D}$ is a symmetric distribution about zero and $\mathbb{P}_{\nu\sim\mathcal{D}}[\nu \in [-1,1]] = 1 - o(N^{-1})$; and*
- *there is a level of signal $\tau_{\mathcal{U}} \in \mathbb{R}$ such that for all $\nu \in [-1,1]$ such that for any fixed constant $K > 0$, it holds that*

$$\left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x)\right| = O_N(\tau_{\mathcal{U}}) \quad \text{and} \quad \left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) - 2\right| = O_N\left(\tau_{\mathcal{U}}^2\right)$$

with probability at least $1 - O\left(N^{-K}\right)$ over each of $\mathcal{P}_\nu, \mathcal{P}_{-\nu}$ and $\mathcal{Q}$.

In our reduction k-BPDS-TO-ISGM, we truncate Gaussians to generate the input distributions Tern. In Figure 16, $\text{TR}_\tau : \mathbb{R} \to \{-1, 0, 1\}$ denotes the truncation map given by

$$\text{TR}_\tau(x) = \begin{cases} 1 & \text{if } x > |\tau| \\ 0 & \text{if } -|\tau| \le x \le |\tau| \\ -1 & \text{if } x < -|\tau| \end{cases}$$

The following simple lemma on truncating symmetric triples of Gaussian distributions will be important in the proofs in this section. Its proof is a direct computation and is deferred to Appendix R.2.

Lemma 103 (Truncating Gaussians) *Let $\tau > 0$ be constant, $\mu > 0$ be tending to zero and let a, μ_1, μ_2 be such that*

$$\begin{aligned} \text{TR}_\tau(\mathcal{N}(\mu, 1)) &\sim \text{Tern}(a, \mu_1, \mu_2) \\ \text{TR}_\tau(\mathcal{N}(-\mu, 1)) &\sim \text{Tern}(a, -\mu_1, \mu_2) \\ \text{TR}_\tau(\mathcal{N}(0, 1)) &\sim \text{Tern}(a, 0, 0) \end{aligned}$$

Then it follows that $a > 0$ is constant, $0 < \mu_1 = \Theta(\mu)$ and $0 < \mu_2 = \Theta(\mu^2)$.

We now will prove our main approximate Markov transition guarantees for k-BPDS-TO-GLSM. The proof follows from combining Theorem 46, Lemma 25 and an application of tensorization of d_{TV}.

Theorem 104 (Reduction from k-BPDS to GLSM) *Let n be a parameter and $w(n) = \omega(1)$ be a slow-growing function. Fix initial and target parameters as follows:*

- Initial k-BPDS Parameters: *vertex counts on each side m and n that are polynomial in one another, dense subgraph dimensions k_m and k_n where k_n divides n, constant edge probabilities $0 < q < p \leq 1$ and a partition F of $[n]$.*
- Target GLSM Parameters: *(N, d) satisfying $wN \leq n$, $N \geq n^{c'}$ for some constant $c' > 0$ and $m \leq d \leq \text{poly}(n)$, target distribution collection $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu \in \mathbb{R}}) \in \text{UC}(N)$ satisyfing that*

$$0 < \tau_{\mathcal{U}} \leq c \cdot \sqrt{\frac{k_n}{n \log n}}$$

for a sufficiently small constant $c > 0$.

Let $\mathcal{A}(M)$ denote k-BPDS-TO-GLSM applied to the adjacency matrix M with these parameters. Then $\mathcal{A}$ runs in $\text{poly}(m, n)$ time and it follows that

$$d_{TV}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S \times T, p, q)\right), \text{GLSM}_D(N, S, d, \mathcal{U})\right) = o(1) + O\left(w^{-1} + k_n^{-2}m^{-2}r^{-2t} + n^{-2} + N^{-3}d^{-3}\right)$$
$$d_{TV}\left(\mathcal{A}\left(\text{Bern}(q)^{\otimes m\times n}\right), \mathcal{Q}^{\otimes d\times N}\right) = O\left(k_n^{-2}m^{-2}r^{-2t} + n^{-2} + N^{-3}d^{-3}\right)$$

for all subsets $S \subseteq [m]$ with $|S| = k_m$ and subsets $T \subseteq [n]$ with $|T| = k_n$ and $|T \cap F_i| = 1$ for each $1 \leq i \leq k_n$.

Proof Let $\mathcal{A}_1$ denote Step 1 of $\mathcal{A}$ with input M and output $(Z_1, Z_2, \ldots, Z_N)$. First note that $2^t = \Theta(n/k_n)$ by the definition of t and $\log m = \Theta(\log n)$ since m and n are polynomial in one another. Thus for a small enough choice of $c_1 > 0$, we have

$$\mu = c_1 \cdot \sqrt{\frac{k_n}{n \log n}} \leq \frac{2^{-(t+1)/2}}{2\sqrt{6\log(k_n m \cdot 2^t) + 2\log(p-q)^{-1}}} \cdot \min\left\{\log\left(\frac{p}{q}\right), \log\left(\frac{1-q}{1-p}\right)\right\}$$

since p and q are constants. Therefore μ satisfies the conditions needed to apply Theorem 46 to $\mathcal{A}_1$. Now let $\mathcal{A}_2$ denote Step 2 of $\mathcal{A}$ with input $(Z_1, Z_2, \ldots, Z_N)$ and output $(X_1, X_2, \ldots, X_N)$. First suppose that $(Z_1, Z_2, \ldots, Z_N) \sim \text{ISGM}_D(N, S, d, \mu, 1/2)$ or in other words where

$$Z_i \sim_{\text{i.i.d.}} \text{MIX}_{1/2}\left(\mathcal{N}(\mu \cdot \mathbf{1}_S, I_d), \mathcal{N}(-\mu \cdot \mathbf{1}_S, I_d)\right)$$

For the next part of this argument, we condition on: (1) the entire vector $\nu = (\nu_1, \nu_2, \ldots, \nu_N)$; and (2) the subset $P \subseteq [N]$ of sample indices corresponding to the positive part $\mathcal{N}(\mu \cdot \mathbf{1}_S, I_d)$ of the mixture. Let $\mathcal{C}(\nu, P)$ denote the event corresponding to this conditioning. After truncating according to TR_τ, by Lemma 103 the resulting entries are distributed as

$$\text{TR}_\tau(Z_{ij}) \sim \begin{cases} \text{Tern}(a, \mu_1, \mu_2) & \text{if } (i,j) \in S \times P \\ \text{Tern}(a, -\mu_1, \mu_2) & \text{if } (i,j) \in S \times P^C \\ \text{Tern}(a, 0, 0) & \text{if } i \notin S \end{cases}$$

Furthermore, these entries are all independent conditioned on (ν, P). Since τ is constant, Lemma 103 also implies that $a \in (0,1)$ is constant, $\mu_1 = \Theta(\mu)$ and $\mu_2 = \Theta(\mu^2)$. Let S_ν be

$$S_\nu = \left\{x \in X : 2|\mu_1| \geq \left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x)\right| \quad \text{and} \quad \frac{2|\mu_2|}{\max\{a, 1-a\}} \geq \left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) - 2\right|\right\}$$

as in Lemma 25. Since $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}) \in \text{UC}(N)$ has level of signal $\tau_{\mathcal{U}} \le c' \cdot \mu$ for a sufficiently small constant $c' > 0$, we have by definition that $\{x \in S_{\nu_i}\}$ occurs with probability at least $1 - \delta_1$ where $\delta_1 = O(n^{-4-K_1})$ over each of $\mathcal{P}_{\nu_i}, \mathcal{P}_{-\nu_i}$ and $\mathcal{Q}$, where $K_1 > 0$ is a constant for which $d = O(n^{K_1})$. Here, we are implicitly using the fact that $N \ge n^{c'}$ for some constant $c' > 0$.

Now consider applying Lemma 25 to each application of 3-SRK in Step 2 of $\mathcal{A}$. Note that $|\mu_1|^{-1} = O(\sqrt{n \log n})$ and $|\mu_2|^{-1} = O(n \log n)$ since $\mu = \Omega(\sqrt{k_n/n \log n})$ and $k_n \ge 1$. Now consider the d-dimensional vectors $X'_1, X'_2, \dots, X'_N$ with independent entries distributed as

$$X'_{ij} \sim \begin{cases} \mathcal{P}_{\nu_i} & \text{if } (i,j) \in S \times P \\ \mathcal{P}_{-\nu_i} & \text{if } (i,j) \in S \times P^C \\ \mathcal{Q} & \text{if } i \notin S \end{cases}$$

The tensorization property of d_{TV} from Fact 15 implies that

$$\begin{aligned}
&d_{\text{TV}}\left(\mathcal{L}(X_1, X_2, \dots, X_N|\nu, P), \mathcal{L}(X'_1, X'_2, \dots, X'_N|\nu, P)\right) \\
&\quad\le \sum_{i=1}^{N}\sum_{j=1}^{d} d_{\text{TV}}\left(\mathcal{L}(X_{ij}|\nu, P), \mathcal{L}(X'_{ij}|\nu, P)\right) \\
&\quad\le \sum_{i=1}^{N}\sum_{j=1}^{d} d_{\text{TV}}\left(3\text{-SRK}(\text{TR}_\tau(Z_{ij}), \mathcal{P}_{\nu_i}, \mathcal{P}_{-\nu_i}, \mathcal{Q}), \mathcal{L}(X'_{ij}|\nu, P)\right) \\
&\quad\le Nd\left[2\delta_1\left(1 + |\mu_1|^{-1} + |\mu_2|^{-1}\right) + \left(\frac{1}{2} + \delta_1\left(1 + |\mu_1|^{-1} + |\mu_2|^{-1}\right)\right)^{N_{\text{it}}}\right] \\
&\quad= O\left(n^{-2} + N^{-3}d^{-3}\right)
\end{aligned}$$

since $N \le n$, $\delta_1 = O(n^{-4}d^{-1})$, $N_{\text{it}} = \lceil 4\log(dN)\rceil$ and by the total variation upper bounds in Lemma 25.

We now will drop the conditioning on (ν, P) and average over $\nu \sim \mathcal{D}'$ and $P \sim \text{Unif}\left[2^{[N]}\right]$. Observe that, when not conditioned on (ν, P), it holds that

$$(X'_1, X'_2, \dots, X'_N) \sim \text{GLSM}_D\left(N, S, d, \left(\mathcal{D}', \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}\right)\right)$$

where $\mathcal{D}'$ is $\mathcal{D}$ conditioned to lie in $[-1, 1]$. Note that here we used the fact that $\mathcal{D}$ and therefore $\mathcal{D}'$ is symmetric about zero. Coupling the latent $\nu_1, \nu_2, \dots, \nu_N$ sampled from $\mathcal{D}$ and $\mathcal{D}'$ and then applying the tensorization property of Fact 15 yields that

$$\begin{aligned}
&d_{\text{TV}}\left(\text{GLSM}_D\left(N, S, d, \left(\mathcal{D}', \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}\right)\right), \text{GLSM}_D\left(N, S, d, (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})\right)\right) \\
&\quad\le d_{\text{TV}}(\mathcal{D}^{\otimes n}, \mathcal{D}'^{\otimes n}) \le N \cdot d_{\text{TV}}(\mathcal{D}, \mathcal{D}') \le N \cdot o(N^{-1}) = o(1)
\end{aligned}$$

where $d_{\text{TV}}(\mathcal{D}, \mathcal{D}') = o(N^{-1})$ follow from the conditioning property of d_{TV} from Fact 15 and the fact that $\mathbb{P}_{\nu\sim\mathcal{D}}[\nu \in [-1, 1]] = 1 - o(N^{-1})$. The triangle inequality and conditioning property of d_{TV} in Fact 15 now imply that

$$\begin{aligned}
&d_{\text{TV}}\left(\mathcal{A}_2\left(\text{ISGM}_D(N, S, d, \mu, 1/2)\right), \text{GLSM}_D\left(N, S, d, \mathcal{U}\right)\right) \\
&\quad\le d_{\text{TV}}\left(\mathcal{L}(X_1, X_2, \dots, X_N), \mathcal{L}(X'_1, X'_2, \dots, X'_N)\right) + d_{\text{TV}}\left(\mathcal{L}(X'_1, X'_2, \dots, X'_N), \text{GLSM}_D\left(N, S, d, \mathcal{U}\right)\right) \\
&\quad\le \mathbb{E}_{\nu\sim\mathcal{D}'}\,\mathbb{E}_{P\sim\text{Unif}\left[2^{[N]}\right]} d_{\text{TV}}\left(\mathcal{L}(X_1, X_2, \dots, X_N|\nu, P), \mathcal{L}(X'_1, X'_2, \dots, X'_N|\nu, P)\right) \\
&\qquad + d_{\text{TV}}\left(\text{GLSM}_D\left(N, S, d, \left(\mathcal{D}', \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}\right)\right), \text{GLSM}_D\left(N, S, d, \mathcal{U}\right)\right) \\
&\quad= o(1) + O\left(n^{-2} + N^{-3}d^{-3}\right)
\end{aligned}$$

Now consider the case when $Z_1, Z_2, \dots, Z_N \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$. Repeating the argument above with $S = \emptyset$ and observing that $(X'_1, X'_2, \dots, X'_N) \sim \mathcal{Q}^{\otimes N}$ yields that

$$d_{\text{TV}}\left(\mathcal{A}_2\left(\mathcal{N}(0, I_d)^{\otimes N}\right), \mathcal{Q}^{\otimes d \times N}\right) = O\left(n^{-2} + N^{-3}d^{-3}\right)$$

We now apply Lemma 16 to the steps $\mathcal{A}_1$ and $\mathcal{A}_2$ under each of H_0 and H_1, as in the proof of Theorem 46. Under H_1, consider Lemma 16 applied to the following sequence of distributions

$$\mathcal{P}_0 = \mathcal{M}_{[m]\times[n]}(S \times T, p, q), \quad \mathcal{P}_1 = \text{ISGM}_D(N, S, d, \mu, 1/2) \quad \text{and} \quad \mathcal{P}_2 = \text{GLSM}_D\left(N, S, d, \mathcal{U}\right)$$

By Theorem 46 and the argument above, we can take

$$\epsilon_1 = O\left(w^{-1} + k_n^{-2}m^{-2}r^{-2t} + n^{-2} + N^{-3}d^{-3}\right) \quad \text{and} \quad \epsilon_2 = o(1) + O\left(n^{-2} + N^{-3}d^{-3}\right)$$

By Lemma 16, we therefore have that

$$d_{\text{TV}}\left(\mathcal{A}\left(\mathcal{M}_{[m]\times[n]}(S \times T, p, q)\right), \text{GLSM}_D(N, S, d, \mathcal{U})\right) = o(1) + O\left(w^{-1} + k_n^{-2}m^{-2}r^{-2t} + n^{-2} + N^{-3}d^{-3}\right)$$

which proves the desired result in the case of H_1. Under H_0, similarly applying Theorem 46, the argument above and Lemma 16 to the distributions

$$\mathcal{P}_0 = \text{Bern}(q)^{\otimes m \times n}, \quad \mathcal{P}_1 = \mathcal{N}(0, I_d)^{\otimes N} \quad \text{and} \quad \mathcal{P}_2 = \mathcal{Q}^{\otimes d \times N}$$

yields the total variation bound

$$d_{\text{TV}}\left(\mathcal{A}\left(\text{Bern}(q)^{\otimes m \times n}\right), \mathcal{Q}^{\otimes d \times N}\right) = O\left(k_n^{-2}m^{-2}r^{-2t} + n^{-2} + N^{-3}d^{-3}\right)$$

which completes the proof of the theorem. ■

We now use this theorem to deduce our universality principle for lower bounds in GLSM. The proof of this next theorem is similar to that of Theorems 4 and 82 and is deferred to Appendix R.2.

Theorem 11 (Computational Lower Bounds for GLSM) *Let n, k and d be polynomial in each other and such that $k = o(\sqrt{d})$. Suppose that the collections of distributions $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})$ is in $\text{UC}(n)$. Then the k-BPC conjecture or k-BPDS conjecture for constant $0 < q < p \le 1$ both imply a computational lower bound for $\text{GLSM}\,(n, k, d, \mathcal{U})$ at all sample complexities $n = \tilde{o}\left(\tau_{\mathcal{U}}^{-4}\right)$.*

O.2. The Universality Class UC(n) and Level of Signal $\tau_{\mathcal{U}}$

The result in Theorem 11 shows universality of the computational sample complexity of $n = \tilde{\Omega}(\tau_{\mathcal{U}}^{-4})$ for learning sparse mixtures under the mild conditions of $\text{UC}(n)$. In this section, we discuss this lower bound, its implications, the universality class $\text{UC}(n)$ and the level of signal $\tau_{\mathcal{U}}$.

Remarks on UC(n) and $\tau_{\mathcal{U}}$. The conditions for $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}) \in \text{UC}(n)$ and the definition of $\tau_{\mathcal{U}}$ have the following two notable properties.

- *They are defined in terms of marginals*: The class $\text{UC}(n)$ and $\tau_{\mathcal{U}}$ are defined entirely in terms of the likelihood ratios $d\mathcal{P}_\nu/d\mathcal{Q}$ between the planted and non-planted marginals. In particular, they are independent of the sparsity level k and other high-dimensional properties of the distribution GLSM constructed from the $\mathcal{P}_\nu$ and $\mathcal{Q}$. Theorem 11 thus establishes a computational lower bound for GLSM at a sample complexity entirely based on properties of the marginals of $\mathcal{P}_\nu$ and $\mathcal{Q}$.

- *Their dependence on n is negligible*: The parameter n only enters the definitions of $\text{UC}(n)$ and $\tau_{\mathcal{U}}$ through requirements on tail probabilities. When the likelihood ratios $d\mathcal{P}_\nu/d\mathcal{Q}$ are relatively concentrated, the dependence of the conditions in $\text{UC}(n)$ and $\tau_{\mathcal{U}}$ on n is nearly negligible. If the ratios $d\mathcal{P}_\nu/d\mathcal{Q}$ are concentrated under $\mathcal{P}_\nu$ and $\mathcal{Q}$ with exponentially decaying tails, then the tail probability bound requirement of $O(n^{-K})$ only appears as a polylog(n) factor in $\tau_{\mathcal{U}}$. This will be the case in the examples that appear later in this section.

$\mathcal{D}$ and Parameterization over $[-1,1]$. $\mathcal{D}$ and the indices of $\mathcal{P}_\nu$ can be reparameterized without changing the underlying problem. The assumption that $\mathcal{D}$ is symmetric and mostly supported on $[-1,1]$ is for notational convenience. As in the case of $\tau_{\mathcal{U}}$ and the examples later in this section, the tail probability requirement of $o(n^{-1})$ for $\mathcal{D}$ only appears as a polylog(n) factor in the computational lower bound of $n = \tilde{\Omega}(\tau_{\mathcal{U}}^{-4})$ if $\mathcal{D}$ is concentrated with exponential tails.

While the output vectors $(X_1, X_2, \ldots, X_N)$ of our reduction k-BPDS-TO-GLSM are independent, their coordinates have dependence induced by the mixture $\mathcal{D}$. The fact that our reduction samples the ν_i implies that if these values were revealed to the algorithm, the problem would still remain hard: an algorithm for the latter could be used together with the reduction to solve k-PC. However, even given the ν_i for the ith sample, our reduction is such that whether the planted marginals in the ith sample are distributed according to $\mathcal{P}_{\nu_i}$ or $\mathcal{P}_{-\nu_i}$ remains unknown to the algorithm. Intuitively, our setup chooses to parameterize the distribution $\mathcal{D}$ over $[-1,1]$ such that the sign ambiguity between $\mathcal{P}_{\nu_i}$ or $\mathcal{P}_{-\nu_i}$ is what is producing hardness below the sample complexity of $n = \tilde{\Omega}(\tau_{\mathcal{U}}^{-4})$.

Implications for Concentrated LLR. We now give several remarks on $\tau_{\mathcal{U}}$ in the case that the log-likelihood ratios (LLR) $\log d\mathcal{P}_\nu/d\mathcal{Q}(x)$ are sufficiently well-concentrated if $x \sim \mathcal{Q}$ or $x \sim \mathcal{P}_\nu$. Suppose that $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}) \in \text{UC}(n)$, fix some arbitrarily large constant $c > 0$ and fix some $\nu \in [-1,1]$. If $S_{\mathcal{Q}}$ is the common support of the $\mathcal{P}_\nu$ and $\mathcal{Q}$, define S to be

$$S = \left\{ x \in S_{\mathcal{Q}} : c \cdot \tau_{\mathcal{U}} \geq \left| \frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) \right| \quad \text{and} \quad c \cdot \tau_{\mathcal{U}}^2 \geq \left| \frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) - 2 \right| \right\}$$

Suppose that $\tau_{\mathcal{U}} = \Omega(n^{-K})$ for some constant $K > 0$ and let c be large enough that S occurs with probability at least $1 - O(n^{-K})$ under each of $\mathcal{P}_\nu, \mathcal{P}_{-\nu}$ and $\mathcal{Q}$. Note that such a constant c is guaranteed by Definition 13. Now observe that

$$\begin{aligned} d_{\text{TV}}\left(\mathcal{P}_\nu, \mathcal{P}_{-\nu}\right) &= \frac{1}{2} \cdot \mathbb{E}_{x\in\mathcal{Q}} \left[\left| \frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) \right| \right] \\ &\leq \frac{1}{2} \cdot \mathbb{E}_{x\in\mathcal{Q}} \left[\left| \frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) \right| \cdot \mathbf{1}_S(x) \right] + \frac{1}{2} \cdot \mathcal{P}_\nu\left[S^C\right] + \frac{1}{2} \cdot \mathcal{P}_{-\nu}\left[S^C\right] \\ &\leq c \cdot \tau_{\mathcal{U}} + O\left(n^{-K}\right) = O\left(\tau_{\mathcal{U}}\right) \end{aligned}$$

A similar calculation with the second condition defining S shows that

$$d_{\text{TV}}\left(\text{MIX}_{1/2}\left(\mathcal{P}_\nu, \mathcal{P}_{-\nu}\right), \mathcal{Q}\right) = O\left(\tau_{\mathcal{U}}^2\right)$$

If the LLRs $\log d\mathcal{P}_\nu/d\mathcal{Q}$ are sufficiently well-concentrated, then the random variables

$$\left| \frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) \right| \quad \text{and} \quad \left| \frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) - 2 \right|$$

will also concentrate around their means if $x \sim \mathcal{Q}$. LLR concentration also implies that this is true if $x \sim \mathcal{P}_\nu$ or $x \sim \mathcal{P}_{-\nu}$. Thus, under sufficient concentration, the definition of the level of signal $\tau_{\mathcal{U}}$ reduces to the much more interpretable pair of upper bounds

$$d_{\mathrm{TV}}\left(\mathcal{P}_\nu, \mathcal{P}_{-\nu}\right) = O\left(\tau_{\mathcal{U}}\right) \quad \text{and} \quad d_{\mathrm{TV}}\left(\mathrm{MIX}_{1/2}\left(\mathcal{P}_\nu, \mathcal{P}_{-\nu}\right), \mathcal{Q}\right) = O\left(\tau_{\mathcal{U}}^2\right)$$

These conditions directly measure the amount of statistical signal present in the planted marginals $\mathcal{P}_\nu$. The relevant calculations for an example application of Theorem 11 when the LLR concentrates is shown below for sparse PCA. In Brennan et al. (2019a), various assumptions of concentration of the LLR and analogous implications for computational lower bounds in submatrix detection are analyzed in detail. We refer the reader to Sections 3 and 9 of Brennan et al. (2019a) for the calculations needed to make the discussion here precise.

We remark that, assuming sufficient concentration on the LLR, the analysis of the k-sparse eigenvalue statistic from Berthet and Rigollet (2013a) yields an information-theoretic upper bound for GLSM. Given GLSM samples $(X_1, X_2, \ldots, X_n)$, consider forming the LLR-processed samples Z_i with

$$Z_{ij} = \mathbb{E}_{\nu \sim \mathcal{D}}\left[\log \frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(X_{ij})\right]$$

for each $i \in [n]$ and $j \in [d]$. Now consider taking the k-sparse eigenvalue of the samples $Z_1, Z_2, \ldots, Z_n$. Under sub-Gaussianity assumptions on the Z_{ij}, the analysis in Theorem 2 of Berthet and Rigollet (2013a) applies. Similarly, the analysis in Theorem 5 of Berthet and Rigollet (2013a) continues to hold, showing that the semidefinite programming algorithm for sparse PCA yields an algorithmic upper bound for GLSM. As information-theoretic limits and algorithms are not the focus of this paper, we omit the technical details needed to make this rigorous.

In many setups captured by GLSM such as sparse PCA, learning sparse mixtures of Gaussians and learning sparse mixtures of Rademachers, these analyses and our lower bound in Theorem 11 together yield a k-to-k^2 statistical-computational gap. How our lower bound yields a k^2 dependence in the computational barriers for these problems is discussed below.

Sparse PCA and Specific Distributions. One specific example captured by our universality principle and that falls under the concentrated LLR setup discussed above is sparse PCA in the spiked covariance model. The statistical-computational gaps of sparse PCA have been characterized based on the planted clique conjecture in a line of work (Berthet and Rigollet, 2013b,a; Wang et al., 2016b; Gao et al., 2017; Brennan et al., 2018; Brennan and Bresler, 2019). We show that our universality principle faithfully recovers the k-to-k^2 gap for sparse PCA shown in Berthet and Rigollet (2013b), Berthet and Rigollet (2013a), Wang et al. (2016b), Gao et al. (2017) and Brennan et al. (2018) assuming the k-BPDS conjecture. As discussed in Section K, also the k-BPC, k-PDS or k-PC conjectures therefore yields nontrivial lower bounds. We remark that Brennan and Bresler (2019) shows stronger hardness based on weaker forms of the PC conjecture.

We show in the next lemma that sparse PCA corresponds to $\mathrm{GLSM}\left(n, k, d, \mathcal{U}\right)$ for a proper choice of $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu \in \mathbb{R}}) \in \mathrm{UC}(n)$ and $\tau_{\mathcal{U}}$ so that the lower bound $n = \tilde{\Omega}(\tau_{\mathcal{U}}^{-4})$ exactly corresponds to the conjectured computational barrier in Sparse PCA. Recall that the hypothesis testing problem $\mathrm{SPCA}(n, k, d, \theta)$ has hypotheses

$$\begin{aligned} H_0 &: (X_1, X_2, \ldots, X_n) \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d) \\ H_1 &: (X_1, X_2, \ldots, X_n) \sim_{\text{i.i.d.}} \mathcal{N}\left(0, I_d + \theta v v^\top\right) \end{aligned}$$

where v is a k-sparse unit vector in $\mathbb{R}^d$ chosen uniformly at random among all such vectors with nonzero entries equal to $1/\sqrt{k}$.

Lemma 105 (Lower Bounds for Sparse PCA) *If, then* $\text{SPCA}(n,k,d,\theta)$ *can be expressed as the instance* $\text{GLSM}(n,k,d,\mathcal{U})$ *where* $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}}) \in \text{UC}(n)$ *is given by*

$$\mathcal{P}_\nu = \mathcal{N}\left(2\nu\sqrt{\frac{\theta\log n}{k}}, 1\right) \text{ for all } \nu\in\mathbb{R}, \quad \mathcal{Q} = \mathcal{N}(0,1) \quad \text{and} \quad \mathcal{D} = \mathcal{N}\left(0, \frac{1}{4\log n}\right)$$

and has valid level of signal $\tau_{\mathcal{U}} = \Theta\left(\sqrt{\frac{\theta(\log n)^2}{k}}\right)$ *if it holds that* $\theta(\log n)^2 = o(k)$.

Proof Note that if $X \sim \mathcal{N}\left(0, I_d + \theta vv^\top\right)$ then X can be written as

$$X = 2\sqrt{\theta\log n}\cdot gv + G \quad \text{where } g\sim\mathcal{N}\left(0,\frac{1}{4\log n}\right) \text{ and } G\sim\mathcal{N}(0,I_d)$$

and where g and G are independent. This follows from the fact that the random variable on the right-hand side above is a jointly Gaussian vector with covariance matrix given by the sum of the covariance matrices of the individual terms. This observation implies that $\text{SPCA}(n,k,d,\theta)$ is exactly the problem $\text{GLSM}(n,k,d,\mathcal{U})$. Now observe that the probability that $x\sim\mathcal{D}$ satisfies $x\in[-1,1]$ is $1-o(n^{-1})$ by standard Gaussian tail bounds. Fix some $\nu\in[-1,1]$ and let $t = 2\nu\sqrt{\frac{\theta\log n}{k}}$. Note that

$$\left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x)\right| = \left|e^{tx-t^2/2} - e^{-tx-t^2/2}\right| = \Theta\left(|tx|\right)$$

if $|tx| = o(1)$. As long as $x = O(\sqrt{\log n})$, it follows that $|tx| = O(\tau_{\mathcal{U}}) = o(1)$ from the definition of $\tau_{\mathcal{U}}$ and fact that $\theta(\log n)^2 = o(k)$. Note that $x = O(\sqrt{\log n})$ occurs with probability at least $1-O(n^{-K})$ for any constant $K>0$ under each of $\mathcal{P}_\nu$ where $\nu\in[-1,1]$ and $\mathcal{Q}$ by standard Gaussian tail bounds. Now observe that

$$\left|\frac{d\mathcal{P}_\nu}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_{-\nu}}{d\mathcal{Q}}(x) - 2\right| = \left|e^{tx-t^2/2} + e^{-tx-t^2/2} - 2\right| = \Theta(t^2)$$

holds if $|tx| = o(1)$, which is true as long as $x = O(\sqrt{\log n})$ and thus holds with probability $1-O(n^{-K})$ for any fixed $K>0$. Since $t^2 = O(\tau_{\mathcal{U}}^2)$ for any $\nu\in[-1,1]$, this completes the proof that $\mathcal{U}\in\text{UC}(n)$ with level of signal $\tau_{\mathcal{U}}$. ■

Combining this lemma with Theorem 11 yields the k-BPDS conjecture implies a computational lower bound for Sparse PCA at the barrier $n = \tilde{o}(k^2/\theta^2)$ as long as $\theta(\log n)^2 = o(k)$ and $k = o(\sqrt{d})$, which matches the planted clique lower bounds in Berthet and Rigollet (2013b), Berthet and Rigollet (2013a), Wang et al. (2016b), Gao et al. (2017) and Brennan et al. (2018). Similar calculations to those in the above corollary can be used to identify the computational lower bound implied by Theorem 11 for many other choices of $\mathcal{U} = (\mathcal{D},\mathcal{Q},\{\mathcal{P}_\nu\}_{\nu\in\mathbb{R}})\in\text{UC}(n)$. Some examples are:

- Balanced sparse Gaussian mixtures where $\mathcal{Q}=\mathcal{N}(0,1)$, $\mathcal{P}_\nu = \mathcal{N}(\theta\nu,1)$ where $\mathcal{D}$ is any symmetric distribution over $[-1,1]$ can be shown to satisfy that $\tau_{\mathcal{U}} = \Theta\left(\theta\sqrt{\log n}\right)$ if $\theta\sqrt{\log n} = o(1)$.

- The Bernoulli case where $\mathcal{Q} = \text{Bern}(1/2)$, $\mathcal{P}_\nu = \text{Bern}(1/2 + \theta\nu)$ and $\mathcal{D}$ is any symmetric distribution over $[-1, 1]$ can be shown to satisfy that $\tau_{\mathcal{U}} = \Theta(\theta)$ if $\theta \leq 1/2$.
- Sparse mixtures of exponential distributions where $\mathcal{Q} = \text{Exp}(\lambda)$, $\mathcal{P}_\nu = \text{Exp}(\lambda + \theta\nu)$ and $\mathcal{D}$ is any symmetric distribution over $[-1, 1]$ can be shown to satisfy that $\tau_{\mathcal{U}} = \tilde{\Theta}\left(\theta\lambda^{-1}\log n\right)$ if it holds that $\theta \log n = o(\lambda)$.
- Sparse mixtures of centered Gaussians with difference variances where $\mathcal{Q} = \mathcal{N}(0, 1)$, $\mathcal{P}_\nu = \mathcal{N}(0, 1 + \theta\nu)$ and $\mathcal{D}$ is any symmetric distribution over $[-1, 1]$ can be shown to satisfy that $\tau_{\mathcal{U}} = \Theta(\theta \log n)$ if $\theta \log n = o(1)$.

We remark that $\tau_{\mathcal{U}}$ can be calculated for many more choices of $\mathcal{D}$, $\mathcal{Q}$ and $\mathcal{P}_\nu$ using the computations outlined in the discussion above on the implications of our result for concentrated LLR.

Appendix P. Computational Lower Bounds for Recovery and Estimation

In this section, we outline several ways to deduce that our reductions to the hypothesis testing formulations in the previous section imply computational lower bounds for natural recovery and estimation formulations of the problems introduced in Section 3. We first introduce a notion of average-case reductions in total variation between recovery problems and note that most of our reductions satisfy these stronger conditions in addition to those in Section E.2. We then discuss alternative methods of obtaining hardness of recovery and estimation in the problems that we consider directly from computational lower bounds for detection.

In the previous section, we showed that lower bounds for our detection formulations of RSME and GLSM directly imply lower bounds for natural estimation and recovery variants, respectively. In Section N, we showed that our lower bounds against blackboxes solving the detection formulation of tensor PCA with a low false positive probability of error directly implies hardness of estimating v in ℓ_2 norm. As discussed in Section B.5, the problems of recovering the hidden partitions in GHPM and BHPM have very different barriers than the testing problem we consider in this work. In this section, we discuss recovery and estimation hardness for the remaining problems from Section 3.

P.1. Our Reductions and Computational Lower Bounds for Recovery

Similar to the framework in Section E.2 for reductions showing hardness of detection, there is a natural notion of a reduction in total variation transferring computational lower bounds between recovery problems. Let $\mathcal{P}(n, \tau)$ denote the recovery problem of estimating $\theta \in \Theta_{\mathcal{P}}$ within some small loss $\ell_{\mathcal{P}}(\theta, \hat{\theta}) \leq \tau$ given an observation from the distribution $\mathcal{P}_D(\theta)$. Here, n is any parameterization such that this observation has size $\text{poly}(n)$ and, as per usual, $\ell_{\mathcal{P}}$, $\Theta_{\mathcal{P}}$ and τ are implicitly functions of n. Define the problem $\mathcal{P}'(N, \tau')$ analogously. The following is the definition of a reduction in total variation between $\mathcal{P}$ and $\mathcal{P}'$.

Definition 106 (Reductions in Total Variation between Recovery Problems) *A $\text{poly}(n)$ time algorithm $\mathcal{A}$ sending valid inputs for $\mathcal{P}(n, \tau)$ to valid inputs for $\mathcal{P}'(N, \tau')$ is a reduction in total variation from $\mathcal{P}$ to $\mathcal{P}'$ if the following criteria are met for all $\theta \in \Theta_{\mathcal{P}}$:*

1. *There is a distribution $\mathcal{D}(\theta)$ over $\Theta_{\mathcal{P}'}$ such that*

$$d_{TV}\left(\mathcal{A}(\mathcal{P}_D(\theta)), \mathbb{E}_{\theta' \sim \mathcal{D}(\theta)} \mathcal{P}'_D(\theta')\right) = o_n(1)$$

2. *There is a* $\mathrm{poly}(n)$ *time randomized algorithm* $\mathcal{B}(X, \hat{\theta}')$ *mapping instances* X *of* $\mathcal{P}(n, \tau)$ *and* $\hat{\theta}' \in \Theta_{\mathcal{P}'}$ *to* $\hat{\theta} \in \Theta_{\mathcal{P}}$ *with the following property: if* $X \sim \mathcal{P}_D(\theta)$*,* θ' *is an arbitrary element of* $\operatorname{supp} \mathcal{D}(\theta)$ *and* $\hat{\theta}'$ *is guaranteed to satisfy that* $\ell_{\mathcal{P}'}(\theta', \hat{\theta}') \le \tau'$*, then* $\mathcal{B}(X, \hat{\theta}')$ *outputs some* $\hat{\theta}$ *with* $\ell_{\mathcal{P}}(\theta, \hat{\theta}) \le \tau$ *with probability* $1 - o_n(1)$*.*

While this definition has a number of technical conditions, it is conceptually simple. A randomized algorithm $\mathcal{A}$ is a reduction in total variation from $\mathcal{P}$ to $\mathcal{P}'$ if it maps a sample from the conditional distribution $\mathcal{P}_D(\theta)$ approximately to a sample from a mixture of $\mathcal{P}_D(\theta')$, where the mixture is over a distribution $\mathcal{D}(\theta)$ determined by θ. Furthermore, there must be an efficient way $\mathcal{B}$ to recover a good estimate $\hat{\theta}$ of θ given a good estimate $\hat{\theta}'$ of θ' and the original instance X of $\mathcal{P}$. The reason that (2) must be true for any $\theta' \in \operatorname{supp} \mathcal{D}(\theta)$ is that, to transfer recovery hardness from $\mathcal{P}$ to $\mathcal{P}'$, the algorithm $\mathcal{B}$ will be applied to the output θ' of a blackbox solving $\mathcal{P}'$ applied to $\mathcal{A}(X)$. In this setting, θ' and X are dependent and allowing $\theta' \in \operatorname{supp} \mathcal{D}(\theta)$ in the definition above accounts for this. Note that, as per usual, $\mathcal{A}$ must satisfy the properties in the definition above oblivious to θ. The following lemma shows that Definition 106 fulfills its objective and transfers hardness of recovery from $\mathcal{P}$ to $\mathcal{P}'$. Its proof is simple and deferred to Appendix Q.1.

Lemma 107 *Suppose that there is reduction* $\mathcal{A}$ *from* $\mathcal{P}(n, \tau)$ *to* $\mathcal{P}'(N, \tau')$ *satisfying the conditions in Definition 106. If there is a polynomial time algorithm* $\mathcal{E}'$ *solving* $\mathcal{P}'(N, \tau')$ *with probability at least* p*, then there is a polynomial time algorithm* $\mathcal{E}$ *solving* $\mathcal{P}(n, \tau)$ *with probability at least* $p - o_n(1)$*.*

The recovery variants of the problems we consider all take the form of $\mathcal{P}(n, \tau)$. For example, $\Theta_{\mathcal{P}}$ is the set of k-sparse vectors of bounded norm and $\ell_{\mathcal{P}}$ is ℓ_2 in MSLR, and $\Theta_{\mathcal{P}}$ is the set of (n/k)-subsets of $[n]$ and $\ell_{\mathcal{P}}$ is the size of the symmetric difference between two (n/k)-subsets in ISBM. In RSLR, $\Theta_{\mathcal{P}}$ can be taken to be the set of al $(u, \mathcal{A})$ where u is a k-sparse vector of bounded norm and $\mathcal{A}$ is a valid adversary. The loss $\ell_{\mathcal{P}}$ is then independent of $\mathcal{A}$ and given by the ℓ_2 norm on u. Throughout Parts II and III, the guarantees we proved for our reductions among the hypothesis testing formulations from Section E.3 generally took the form of condition (1) in Definition 106. Some reductions had a post-processing step where coordinates in the output instance are randomly permuted or subsampled, but these can simply be removed to yield a guarantee matching the form of (1). In light of this and Lemma 107, it suffices to show that our reductions also satisfy condition (2) in Definition 106. We outline how to construct these algorithms $\mathcal{B}$ for each of our remaining problems below.

Reductions from BPC and k-BPC. All of our reductions from BPC and k-BPC to RSME, NEG-SPCA, MSLR and RSLR map from an instance with left biclique vertex set S with $|S| = k_m$ to an instance with hidden vector $u = \gamma \cdot k_m^{-1/2} \cdot \mathbf{1}_S$ for some $\gamma \in (0, 1)$. In the notation of Definition 106, $\mathcal{D}(S)$ is a point mass on u. We now outline how such reductions imply hardness of estimation up to any ℓ_2 error $\tau' = o(\gamma)$.

To verify condition (2) of Definition 106, it suffices to give an efficient algorithm $\mathcal{B}$ recovering S and the right biclique vertices S' from the original BPC or k-BPC instance G and an estimate $\hat{u}$ satisfying that $\|\hat{u} - \gamma \cdot k_m^{-1/2} \cdot \mathbf{1}_S\|_2 \le \tau'$. Suppose that $|S| = k_m$ and $|S'|$ are both $\omega(\log n)$. Let $\hat{S}$ be the set of the largest k_m entries of $\hat{u}$ and note that $\|\gamma^{-1} \cdot \hat{u} - k_m^{-1/2} \cdot \mathbf{1}_S\|_2 = o(1)$, which can be verified to imply that at least $(1 - o(1))k_m$ of $\hat{S}$ must be in S. A union bound and Chernoff bound can be used to show that, in a BPC instance with left and right biclique sets S and S', there is no

right vertex in $[n]\backslash S'$ with at least $3k_m/4$ neighbors in S with probability $1 - o_n(1)$ if $k_m \gg \log n$. Therefore S' is exactly the set of right vertices with at least $5k_m/6$ neighbors in $\hat{S}$ with probability $1 - o_n(1)$. Taking the common neighbors of S' now recovers S with high probability. Thus this procedure of taking the k_m largest entries $\hat{S}$ of $\hat{u}$, taking right vertices with many neighbors in $\hat{S}$ and then taking their common neighborhoods, exactly solves the BPC and k-BPC recovery problems. We remark that hardness for these exact recovery problems follows from detection hardness, as bipartite Erdős-Rényi random graphs do not contain bicliques of left and right sizes $\omega(\log n)$ with probability $1 - o_n(1)$.

We remark that for the values of γ in our reductions, the condition $\tau = o(\gamma)$ implies tight computational lower bounds for estimation in RSME, NEG-SPCA, MSLR and RSLR. In particular, for RSME, MSLR and RSLR, we may take τ' to be arbitrarily close to τ in our detection lower bound as long as $\tau' = o(\tau)$. For NEG-SPCA, a natural estimation analogue is to estimate some k-sparse v within ℓ_2 norm τ' given n i.i.d. samples from $\mathcal{N}(0, I_d + vv^\top)$. For this estimation formulation, we may take $\tau' = o(\sqrt{\theta})$ where θ is as in our detection lower bound.

Reductions from k-PC. We now outline how to construct such an algorithm $\mathcal{B}$ for ISBM. We only sketch the details of this construction as a more direct and simpler way to deduce hardness of recovery for ISBM will be discussed in the next section. We remark that a similar construction of $\mathcal{B}$ also verifies condition (2) for our reduction to SEMI-CR.

For simplicity, first consider k-PDS-TO-ISBM without the initial TO-k-PARTITE-SUBMATRIX step and the random permutations of vertex labels in Steps 2 and 4. Let $S \subseteq [kr^t]$ be the vertex set of the planted dense subgraph in M_{PD2} and let F' and F'' be the given partitions of the indices $[kr^t]$ of M_{PD2} and the vertices $[kr\ell]$ of the output graph, respectively. Lemma 86 shows that the output instance of ISBM has its smaller hidden community C_1 of size $k\ell$ on the vertices corresponding to the negative entries of the vector $v_{S,F',F''}(K_{r,t})$. Note that, as a function of this set S, the mixture distribution $\mathcal{D}(S)$ is again a point mass. We now will outline how to approximately recover S given a close estimate $\hat{C}_1$ of C_1. Suppose that $\hat{C}_1$ is a $k\ell$-subset of $[kr\ell]$ such that $|C_1 \cap \hat{C}_1| \geq (1-o(1))k\ell$. Construct the vector $\hat{v}$ given by

$$\hat{v}_i = \frac{1}{\sqrt{r^t(r-1)}} \cdot \begin{cases} 1 & \text{if } i \notin \hat{C}_1 \\ 1-r & \text{if } i \in \hat{C}_1 \end{cases}$$

Since $\ell = \Theta(r^{t-1})$, a direct calculation shows that $\|\hat{v} - v_{S,F',F''}(K_{r,t})\|_2 = o(\sqrt{k})$. For each part F_i'', consider the vector in $\mathbb{R}^{r\ell}$ formed by restricting $\hat{v}$ to the indices in F_i'' and identifying these indices with $[r\ell]$ in increasing order. For each such vector, find the closest column of $K_{r,t}$ to this vector in ℓ_2 norm. If the index of this column is j, add the jth smallest element of F_i' to $\hat{S}$. We claim that the resulting set $\hat{S}$ contains at least $(1 - o(1))k$ elements of S. The singular values of $K_{r,t}$ computed in Lemma 30 can be used to show that any two columns of $K_{r,t}$ are separated by an ℓ_2 distance of $\Omega(1)$. Any part F_i' for which the correct $j \in S \cap F_i'$ was not added to $\hat{S}$ must have satisfied that $\hat{v}$ restricted to the part F_i'' was an ℓ_2 distance of $\Omega(1)$ from the corresponding restriction of $v_{S,F',F''}(K_{r,t})$. Since $\|\hat{v} - v_{S,F',F''}(K_{r,t})\|_2 = o(\sqrt{k})$, the number of such j incorrectly added to $\hat{S}$ is $o(k)$, verifying the claim.

Now consider k-PDS-TO-ISBM with its first step and the random permutations. Since the random index permutation in TO-k-PARTITE-SUBMATRIX and the subsequent random permutations in Steps 2 and 4 are all generated by the reduction, they can also be remembered and used in the algorithm $\mathcal{B}$ recovering the clique of the input k-PC instance. When combined with the subroutine

recovering $\hat{S}$ from $\hat{C}_1$, these permutations are sufficient to identify a set of k vertices overlapping with the clique in at least $(1-o(1))k$ vertices. Now using a similar procedure to the one mentioned above for BPC, together with the input k-PC instance G, this is sufficient to exactly recover the hidden clique vertices.

P.2. Relationship Between Detection and Recovery

As shown in the previous section, computational lower bounds from recovery can generally be deduced from our reductions because they are also reductions in total variation between recovery problems. We now will outline how our computational lower bounds for detection all either directly or almost directly imply hardness of recovery. As in Section 10 of Brennan et al. (2018), our approach is to produce two independent instances X and X' from $\mathcal{P}_D(\theta)$ without knowing θ, to use X to recover an estimate $\hat{\theta}$ of θ and then to verify that $\hat{\theta}$ is a good estimate of θ using X'. If $\hat{\theta}$ is confirmed to closely approximate θ using X', then output H_1, and otherwise output H_0. This recipe shows detection is easier than recovery as long as there are efficient ways to produce the pair (X, X') and to verify $\hat{\theta}$ is a good estimate given a fresh sample X'. In general, the purpose of cloning into the pair (X, X') is to sidestep the fact that X and $\hat{\theta}$ are dependent random variables, which complicates analyzing the verification step. In contrast, $\hat{\theta}$ and X' are conditionally independent given θ. We now show that this recipe applies to each of our problems.

Sample Splitting. In problems with samples, a natural way to produce X and X' is to simply split the set of samples into two groups. This yields a means to directly transfer computational lower bounds from detection to recovery for RSME, NEG-SPCA, MSLR and RSLR. As we already discussed one way our reductions imply computational lower bounds for the recovery variants of these problems in the previous section, we only sketch the main ideas here.

We first show an efficient algorithm for recovery in MSLR yields an efficient algorithm for detection. Consider the detection problem $\text{MSLR}(2n, k, d, \tau)$, and assume there is a blackbox $\mathcal{E}$ solving the recovery problem $\text{MSLR}(n, k, d, \tau')$ with probability $1 - o_n(1)$ for some $\tau' = o(\tau)$. If the samples from $\text{MSLR}(2n, k, d, \tau)$ are $(X_1, y_1), (X_2, y_2), \ldots, (X_{2n}, y_{2n})$, apply $\mathcal{E}$ to $(X_1, y_1), \ldots, (X_n, y_n)$ to produce an estimate $\hat{u}$. Under H_1, there is some true $u = \tau \cdot k^{-1/2} \cdot \mathbf{1}_S$ for some k-set S and it holds that $\|\hat{u} - u\|_2 = o(\tau)$. As in the previous section, taking the largest k coordinates of $\hat{u}$ yields a set $\hat{S}$ containing at least $(1-o(1))k$ elements of S. The idea is now that we almost know the true set S, detection using the second group of n samples essentially reduces to MSLR without sparsity and is easy down to the information-theoretic limit. More precisely, consider using the second half of the samples to form the statistic

$$Z = \frac{1}{\tau^2(1+\tau^2)} \sum_{i=n+1}^{2n} \left(y_i^2 - 1 - \tau^2\right) \cdot \left\langle (X_i)_{\hat{S}}, \hat{u}_{\hat{S}} \right\rangle^2$$

where $v_{\hat{S}}$ denotes the vector equal to v on the indices in $\hat{S}$ and zero elsewhere. Note that conditioned on S, the second group of n samples is independent of $\hat{S}$. Under H_0, it can be verified that $\mathbb{E}[Z] = 0$ and $\text{Var}[Z] = O(n)$. Under H_1, it can be verified that $\|\hat{u}\|_2$ and $\|\hat{u}_{\hat{S}}\|_2$ are both $(1+o(1))\tau$ and furthermore that $\langle u, \hat{u}_{\hat{S}} \rangle \geq (1-o(1))\tau^2$. Now note that since $y_i = R_i \cdot \langle X_i, u \rangle + g_i$ where $g_i \sim \mathcal{N}(0,1)$ and $R_i \sim \text{Rad}$, we have that

$$\left(y_i^2 - 1 - \tau^2\right) \cdot \left\langle (X_i)_{\hat{S}}, \hat{u}_{\hat{S}} \right\rangle^2 = \langle X_i, u \rangle^2 \cdot \left\langle X_i, \hat{u}_{\hat{S}} \right\rangle^2 - \tau^2 \cdot \left\langle X_i, \hat{u}_{\hat{S}} \right\rangle^2 + 2R_i g_i \cdot \langle X_i, u \rangle \cdot \left\langle X_i, \hat{u}_{\hat{S}} \right\rangle^2 + (g_i^2 - 1) \cdot \left\langle X_i, \hat{u}_{\hat{S}} \right\rangle^2$$

The last two terms are mean zero and the second term has expectation $-(1+o(1))\tau^4$ since $\|\hat{u}_{\hat{S}}\|_2 = (1+o(1))\tau$. Directly expanding the first term in terms of the components of X_i yields that its expectation is given by $2\langle u, \hat{u}_{\hat{S}}\rangle^2 + \|u\|_2^2 \cdot \|\hat{u}_{\hat{S}}\|_2^2 \geq 3(1-o(1))\tau^4$. Combining these computations yields that $\mathbb{E}[Z] \geq 2n(1-o(1))\tau^2$, and it can again be verified that $\mathrm{Var}[Z] = O(n)$. Chebyshev's inequality now yields that thresholding Z at $n\tau^2$ distinguishes H_0 and H_1 as long as $\tau^2\sqrt{n} \gg 1$. Since the information-theoretic limit of the detection formulation of MSLR is when $n = \Theta(k \log d/\tau^4)$ (Fan et al., 2018), whenever this problem is possible it holds that $\tau^2\sqrt{n} \gg 1$. Therefore, whenever detection is possible, the reduction outlined above shows how to produce a test solving detection in MSLR using an estimator with ℓ_2 error $\tau' = o(\tau)$.

Similar reductions transfer hardness of recovery to detection for NEG-SPCA, RSME and RSLR. For NEG-SPCA and RSME, the same argument as above can be shown to work with the test statistic given by $Z = \sum_{i=1}^{2n} \langle X_i, \hat{u}_{\hat{S}}\rangle^2$, and the same Z used above for MSLR suffices in the case of RSLR. We remark that to show these statistics Z solve the detection variants of RSME and RSLR, it is important to use detection formulations incorporating the exact form of our adversarial constructions, which are ISGM in the case of RSME and the adversary described in Section I in the case of RSLR. An arbitrary adversary could corrupt instances of RSME and RSLR to cause these statistics Z to not distinguish between H_0 and H_1. Because our detection lower bounds apply to these fixed adversaries rather than requiring an arbitrary adversary, this argument yields the desired hardness of estimation for RSME and RSLR.

Post-Reduction Cloning. In problems without samples, producing the pair (X, X') requires an additional reduction step. We now outline how to produce such a pair and verification step for ISBM. The high-level idea is to stop our reduction to ISBM before the final thresholding step, apply Gaussian cloning as in Section 10 of Brennan et al. (2018), then to continue the reduction with both copies, eventually using one to verify the output of a recovery blackbox applied to the other. A similar argument can be used to show computational lower bounds for recovery in SEMI-CR.

Consider the reduction k-PDS-TO-ISBM without the final thresholding step, outputting the matrix $M_{\mathrm{R}} \in \mathbb{R}^{kr\ell \times kr\ell}$ at the end of Step 3. Now consider adding the following three steps to this reduction, given access to a recovery blackbox $\mathcal{E}$. More precisely, given an instance of $\mathrm{ISBM}(n, k, P_{11}, P_{12}, P_{22})$ with

$$P_{11} = P_0 + \gamma, \quad P_{12} = P_0 - \frac{\gamma}{k-1} \quad \text{and} \quad P_{22} = P_0 + \frac{\gamma}{(k-1)^2}$$

as in Section M.1, suppose $\mathcal{E}$ is guaranteed to output an (n/k)-subset of vertices $\hat{C}_1 \subseteq [n]$ with $|C_1 \cap \hat{C}_1| \geq (1+\epsilon)n/k^2$ with probability $1 - o_n(1)$ for some $\epsilon = \Omega(1)$. Here, C_1 is the true hidden smaller community of the input ISBM instance. Observe that when $\epsilon = \Theta(1)$, the blackbox $\mathcal{E}$ has the weak guarantee of recovering marginally more than a trivial $1/k$ fraction of C_1. This exactly matches the notion of weak recovery discussed in Section B.4.

1. Sample $W \sim \mathcal{N}(0,1)^{\otimes n \times n}$ and form

$$M_{\mathrm{R}}^1 = \frac{1}{\sqrt{2}}(M_{\mathrm{R}} + W) \quad \text{and} \quad M_{\mathrm{R}}^2 = \frac{1}{\sqrt{2}}(M_{\mathrm{R}} - W)$$

2. Using each of M_{R}^1 and M_{R}^2, complete the reduction k-PDS-TO-ISBM omitting the random permutation in Step 4, and complete the additional steps from Corollary 88 replacing μ with $\mu/\sqrt{2}$. Let the two output graphs be G^1 and G^2.

3. Let $\hat{C}_1$ be the output of $\mathcal{E}$ applied to G^1. Output H_0 if the subgraph of G^2 restricted to $\hat{C}_1$ has at least M edges, and output H_1 otherwise.

We now outline how this solves the detection variant of ISBM. Let C_1 be the true hidden smaller community of the instance that k-PDS-TO-ISBM would produce if completed using M_{R}. We claim that G^1 and G^2 are $o(1)$ total variation from independent copies of $\mathrm{ISBM}(n, C_1, P_{11}, P_{12}, P_{22})$ where P_{11}, P_{12} and P_{22} are as above and γ is as in Corollary 88, but defined using $\mu/\sqrt{2}$ instead of μ. To see this, note that M_{R} is $o(1)$ total variation from the distribution

$$M'_{\mathrm{R}} = \frac{\mu(r-1)}{r} \cdot v(C_1)v(C_1)^\top + Y \quad \text{where} \quad v(C_1)_i = \frac{1}{\sqrt{r^t(r-1)}} \cdot \begin{cases} 1 & \text{if } i \notin C_1 \\ 1-r & \text{if } i \in C_1 \end{cases}$$

by Lemma 86, where $Y \sim \mathcal{N}(0,1)^{\otimes n \times n}$ and t is the internal parameter used in k-PDS-TO-ISBM. Now it follows that M^1_{R} and M^2_{R} are respectively $o(1)$ total variation from

$$\left(M^1_{\mathrm{R}}\right)' = \frac{\mu(r-1)}{r\sqrt{2}} \cdot v(C_1)v(C_1)^\top + \frac{1}{\sqrt{2}}(Y+W) \quad \text{and}$$
$$\left(M^2_{\mathrm{R}}\right)' = \frac{\mu(r-1)}{r\sqrt{2}} \cdot v(C_1)v(C_1)^\top + \frac{1}{\sqrt{2}}(Y-W)$$

The entries of $\frac{1}{\sqrt{2}}(Y+W)$ and $\frac{1}{\sqrt{2}}(Y-W)$ are all jointly Gaussian and have variance 1. Furthermore, they can all be verified to be uncorrelated, implying that these two matrices are independent copies of $\mathcal{N}(0,1)^{\otimes n \times n}$ and thus $\left(M^1_{\mathrm{R}}\right)'$ and $\left(M^2_{\mathrm{R}}\right)'$ are independent conditioned on C_1. Note that μ has essentially been scaled down by a factor of $\sqrt{2}$ in both of these instances as well. Thus Step 2 above ensures that G^1 and G^2 are $o(1)$ total variation from independent copies of $\mathrm{ISBM}(n, C_1, P_{11}, P_{12}, P_{22})$.

Now consider Step 3 above applied to two exact independent copies of $\mathrm{ISBM}(n, C_1, P_{11}, P_{12}, P_{22})$. The guarantee for $\mathcal{E}$ ensures that $|C_1 \cap \hat{C}_1| \geq (1+\epsilon)n/k^2$ with probability $1 - o_n(1)$. The variance of the number of edges in the subgraph of G^2 restricted to $\hat{C}_1$ is $O(n^2/k^2)$ under both H_0 and H_1, and the expected number of edges in this subgraph is $P_0\binom{n/k}{2}$ under H_0. Under H_1, the expected number of edges is

$$\begin{aligned}
\mathbb{E}\left[|E(G[\hat{C}_1])|\right] &= \binom{|C_1 \cap \hat{C}_1|}{2} P_{11} + |C_1 \cap \hat{C}_1| \cdot \left(\frac{n}{k} - |C_1 \cap \hat{C}_1|\right) P_{12} + \binom{\frac{n}{k} - |C_1 \cap \hat{C}_1|}{2} P_{22} \\
&= P_0 \binom{n/k}{2} + \frac{\gamma}{2(k-1)^2} \cdot \left(k|C_1 \cap \hat{C}_1| - \frac{n}{k}\right)^2 - \frac{\gamma}{2(k-1)} \cdot \left((k-2) \cdot |C_1 \cap \hat{C}_1| + \frac{n}{k}\right) \\
&= P_0 \binom{n/k}{2} + \Omega\left(\frac{\gamma \epsilon^2 n^2}{k^4}\right)
\end{aligned}$$

where the last bound holds since $\epsilon = \Omega(1)$ and $k^2 \ll n$.

By Chebyshev's inequality, Step 3 solves the hypothesis testing problem exactly when this difference $\Omega(\gamma\epsilon^2 n^2/k^4)$ grows faster than the $O(n/k)$ standard deviations in the number of edges in the subgraph under H_0 and H_1. This implies that Step 3 succeeds if it holds that $\gamma\epsilon^2 \gg k^3/n$. The Kesten-Stigum threshold corresponds to $\gamma^2 = \tilde{\Theta}(k^2/n)$ and therefore as long as $\epsilon^4 n = \tilde{\omega}(k^4)$, this argument solves the detection problem just below the Kesten-Stigum threshold. When $\epsilon = \Theta(1)$, this argument shows a computational lower bound up to the Kesten-Stigum threshold for weak recovery in ISBM. Since $k^2 = o(n)$ is always true in our formulation of ISBM, setting $\epsilon = \Theta(\sqrt{k})$

yields that for all k it is hard to recover a $\Theta(1/\sqrt{k})$ fraction of the hidden community C_1. This guarantee is much stronger than the analysis in the previous section, which only showed hardness for a blackbox recovering a $1 - o(1)$ fraction of the hidden community. We remark that the same trick used in Step 1 above to produce two independent copies of a matrix with Gaussian noise was used to show estimation lower bounds for tensor PCA in Section N.

Pre-Reduction Cloning. We remark that there is a general alternative method to obtain the pairs (X, X') in our reductions that we sketch here. Consider applying Bernoulli cloning either directly to the input PC or PDS instance or to the output of TO-k-PARTITE-SUBMATRIX, in the case of reductions from k-PC, and then running the remaining parts of our reductions on each of the two resulting copies. Ignoring post-processing steps where we permute vertex labels or subsample the output instance, this general approach can be used to yield two copies of the outputs of our reductions that have the same hidden structure and are conditionally independent given this hidden structure. The same verification steps outlined above can then be applied to obtain our computational lower bounds for recovery.

Part IV
Deferred Proofs

Appendix Q. Deferred Proofs from Part II

Q.1. Proofs of Total Variation Properties

In this section, we present the deferred proofs from Sections E.2 and P. We first prove Lemma 16.

Proof (of Lemma 16) This follows from a simple induction on m. Note that the case when $m = 1$ follows by definition. Now observe that by the data-processing and triangle inequalities of total variation, we have that if $\mathcal{B} = \mathcal{A}_{m-1} \circ \mathcal{A}_{m-2} \circ \cdots \circ \mathcal{A}_1$ then

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathcal{A}(\mathcal{P}_0), \mathcal{P}_m\right) &\le d_{\mathrm{TV}}\left(\mathcal{A}_m \circ \mathcal{B}(\mathcal{P}_0), \mathcal{A}_m(\mathcal{P}_{m-1})\right) + d_{\mathrm{TV}}\left(\mathcal{A}_m(\mathcal{P}_{m-1}), \mathcal{P}_m\right) \\ &\le d_{\mathrm{TV}}\left(\mathcal{B}(\mathcal{P}_0), \mathcal{P}_{m-1}\right) + \epsilon_m \\ &\le \sum_{i=1}^{m} \epsilon_i \end{aligned}$$

where the last inequality follows from the induction hypothesis applied with $m - 1$ to $\mathcal{B}$. This completes the induction and proves the lemma. ■

We now prove Lemma 17 upper bounding the total variation distance between vectors of unplanted and planted samples from binomial distributions.

Proof (of Lemma 17) Given some $P \in [0, 1]$, we begin by computing the χ^2-divergence given by $\chi^2\left(\mathrm{Bern}(P) + \mathrm{Bin}(m-1, Q), \mathrm{Bin}(m, Q)\right)$. For notational convenience, let $\binom{a}{b} = 0$ if $b > a$ or $b < 0$. It follows that

$$1 + \chi^2\left(\mathrm{Bern}(P) + \mathrm{Bin}(m-1, Q), \mathrm{Bin}(m, Q)\right)$$

$$
\begin{aligned}
&= \sum_{t=0}^{m} \frac{\left((1-P)\cdot\binom{m-1}{t}Q^t(1-Q)^{m-1-t} + P\cdot\binom{m-1}{t-1}Q^{t-1}(1-Q)^{m-t}\right)^2}{\binom{m}{t}Q^t(1-Q)^{m-t}} \\
&= \sum_{t=0}^{m}\binom{m}{t}Q^t(1-Q)^{m-t}\left(\frac{m-t}{m}\cdot\frac{1-P}{1-Q}+\frac{t}{m}\cdot\frac{P}{Q}\right)^2 \\
&= \mathbb{E}\left[\left(\frac{m-X}{m}\cdot\frac{1-P}{1-Q}+\frac{X}{m}\cdot\frac{P}{Q}\right)^2\right] \\
&= \mathbb{E}\left[\left(1+\frac{X-mQ}{m}\cdot\frac{P-Q}{Q(1-Q)}\right)^2\right] \\
&= 1+\frac{2(P-Q)}{mQ(1-Q)}\cdot\mathbb{E}[X-mQ]+\frac{(P-Q)^2}{m^2Q^2(1-Q)^2}\cdot\mathbb{E}\left[(X-Qm)^2\right] \\
&= 1+\frac{(P-Q)^2}{mQ(1-Q)}
\end{aligned}
$$

where $X \sim \text{Bin}(m,Q)$ and the second last equality follows from $\mathbb{E}[X] = Qm$ and $\mathbb{E}[(X-Qm)^2] = \text{Var}[X] = Q(1-Q)m$. The concavity of $\log$ implies that $d_{\text{KL}}(\mathcal{P},\mathcal{Q}) \le \log\left(1+\chi^2(\mathcal{P},\mathcal{Q})\right) \le \chi^2(\mathcal{P},\mathcal{Q})$ for any two distributions with $\mathcal{P}$ absolutely continuous with respect to $\mathcal{Q}$. Pinsker's inequality and tensorization of d_{KL} now imply that

$$
\begin{aligned}
&2\cdot d_{\text{TV}}\left(\otimes_{i=1}^{k}\left(\text{Bern}(P_i)+\text{Bin}(m-1,Q)\right),\text{Bin}(m,Q)^{\otimes k}\right)^2 \\
&\quad\le d_{\text{KL}}\left(\otimes_{i=1}^{k}\left(\text{Bern}(P_i)+\text{Bin}(m-1,Q)\right),\text{Bin}(m,Q)^{\otimes k}\right) \\
&\quad= \sum_{i=1}^{k} d_{\text{KL}}\left(\text{Bern}(P_i)+\text{Bin}(m-1,Q),\text{Bin}(m,Q)\right) \\
&\quad\le \sum_{i=1}^{k}\chi^2\left(\text{Bern}(P_i)+\text{Bin}(m-1,Q),\text{Bin}(m,Q)\right) = \sum_{i=1}^{k}\frac{(P_i-Q)^2}{mQ(1-Q)}
\end{aligned}
$$

which completes the proof of the lemma. ■

We now prove Lemma 18 on the total variation distance between two binomial distributions.

Proof (of Lemma 18) By applying the data processing inequality for d_{TV} to the function taking the sum of the coordinates of a vector, we have that

$$
\begin{aligned}
2\cdot d_{\text{TV}}\left(\text{Bin}(n,P),\text{Bin}(n,Q)\right)^2 &\le 2\cdot d_{\text{TV}}\left(\text{Bern}(P)^{\otimes n},\text{Bern}(Q)^{\otimes n}\right)^2 \\
&\le d_{\text{KL}}\left(\text{Bern}(P)^{\otimes n},\text{Bern}(Q)^{\otimes n}\right) \\
&= n\cdot d_{\text{KL}}\left(\text{Bern}(P),\text{Bern}(Q)\right) \\
&\le n\cdot\chi^2\left(\text{Bern}(P),\text{Bern}(Q)\right) \\
&= n\cdot\frac{(P-Q)^2}{Q(1-Q)}
\end{aligned}
$$

The second inequality is an application of Pinsker's, the first equality is tensorization of d_{KL} and the third inequality is the fact that χ^2 upper bounds d_{KL} by the concavity of $\log$. This completes the proof of the lemma. ■

We conclude this section with a proof of Lemma 107, establishing the key property of reductions in total variation among recovery problems.

Proof (of Lemma 107) As in the proof of Lemma 14 from Brennan et al. (2018), this lemma follows from a simple application of the definition of d_{TV}. Suppose that there is such an $\mathcal{E}'$. Now consider the algorithm $\mathcal{E}$ that proceeds as follows on an input X of $\mathcal{P}(n, \tau)$:

1. compute $\mathcal{A}(X)$ and the output $\hat{\theta}'$ of $\mathcal{E}'$ on input $\mathcal{A}(X)$; and
2. output the result $\hat{\theta} \leftarrow \mathcal{B}(X, \hat{\theta}')$.

Suppose that $X \sim \mathcal{P}_D(\theta)$ for some $\theta \in \Theta_{\mathcal{P}}$. Consider a coupling of X, the randomness of $\mathcal{A}$ and $Y \sim \mathbb{E}_{\theta' \sim \mathcal{D}(\theta)} \mathcal{P}'_D(\theta')$ such that $\mathbb{P}[\mathcal{A}(X) \neq Y] = o_n(1)$. Since Y is distributed as a mixture of $\mathcal{P}'_D(\theta')$, conditioned on θ', it holds that $\mathcal{E}'$ succeeds with probability

$$\mathbb{P}\left[\ell_{\mathcal{P}'}(\mathcal{E}'(Y), \theta') \leq \tau' \,\middle|\, \theta'\right] \geq p$$

Marginalizing this over θ' yields that $\mathbb{P}\left[\ell_{\mathcal{P}'}(\mathcal{E}'(Y), \theta') \leq \tau' \text{ for some } \theta' \in \operatorname{supp} \mathcal{D}(\theta)\right] \geq p$. Now since $\mathcal{A}(X) = Y$ is a probability $1 - o_n(1)$ event, we have that the intersection of this and the event above occurs with probability $p - o_n(1)$. Therefore

$$\mathbb{P}\left[\ell_{\mathcal{P}'}(\theta', \hat{\theta}') \leq \tau' \text{ for some } \theta' \in \operatorname{supp} \mathcal{D}(\theta)\right] \geq \mathbb{P}\left[\mathcal{A}(X) = Y \text{ and } \mathcal{E}' \text{ succeeds}\right] \geq p - o_n(1)$$

Now note that the definition of $\mathcal{B}$ implies that

$$\begin{aligned}
\mathbb{P}\left[\ell_{\mathcal{P}}(\theta, \hat{\theta}) \leq \tau\right] &\geq \mathbb{P}\left[\ell_{\mathcal{P}'}(\theta', \hat{\theta}') \leq \tau' \text{ for some } \theta' \in \operatorname{supp} \mathcal{D}(\theta) \text{ and } \mathcal{B} \text{ succeeds}\right] \\
&\geq \mathbb{P}\left[\ell_{\mathcal{P}'}(\theta', \hat{\theta}') \leq \tau' \text{ for some } \theta' \in \operatorname{supp} \mathcal{D}(\theta)\right] - \mathbb{P}\left[\mathcal{B} \text{ fails}\right] \\
&\geq p - o_n(1)
\end{aligned}$$

which completes the proof of the lemma. ■

Q.2. Proofs for To-k-Partite-Submatrix

In this section, we prove Lemma 23, which establishes the approximate Markov transition properties of the reduction TO-k-PARTITE-SUBMATRIX. We first establish analogue of Lemma 6.4 from Brennan et al. (2019a) in the k-partite case to analyze the planted diagonal entries in Step 2 of TO-k-PARTITE-SUBMATRIX.

Lemma 108 (Planting k-Partite Diagonals) *Suppose that $0 < Q < P \leq 1$ and $n \geq \left(\frac{P}{Q} + 1\right) N$ is such that both N and n are divisible by k and $k \leq QN/4$. Suppose that for each $t \in [k]$,*

$$z_1^t \sim \mathrm{Bern}(P), \quad z_2^t \sim \mathrm{Bin}(N/k - 1, P) \quad \text{and} \quad z_3^t \sim \mathrm{Bin}(n/k, Q)$$

are independent. If $z_4^t = \max\{z_3^t - z_1^t - z_2^t, 0\}$*, then it follows that*

$$d_{TV}\left(\otimes_{t=1}^k \mathcal{L}(z_1^t, z_2^t + z_4^t), (\text{Bern}(P) \otimes \text{Bin}(n/k-1, Q))^{\otimes k}\right) \leq 4k \cdot \exp\left(-\frac{Q^2N^2}{48Pkn}\right) + \sqrt{\frac{C_Q k^2}{2n}}$$
$$d_{TV}\left(\otimes_{t=1}^k \mathcal{L}(z_1^t + z_2^t + z_4^t), \text{Bin}(n/k, Q)^{\otimes k}\right) \leq 4k \cdot \exp\left(-\frac{Q^2N^2}{48Pkn}\right)$$

where $C_Q = \max\left\{\frac{Q}{1-Q}, \frac{1-Q}{Q}\right\}$.

Proof Throughout this argument, let v denote a vector in $\{0,1\}^k$. Now define the event

$$\mathcal{E} = \bigcap_{t=1}^k \{z_3^t = z_1^t + z_2^t + z_4^t\}$$

Now observe that if $z_3^t \geq Qn/k - QN/2k + 1$ and $z_2^t \leq P(N/k-1) + QN/2k$ then it follows that $z_3^t \geq 1 + z_2^t \geq v_t + z_2^t$ for any $v_t \in \{0,1\}$ since $Qn \geq (P+Q)N$. Now union bounding the probability that $\mathcal{E}$ does not hold conditioned on z_1 yields that

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}^C \middle| z_1 = v\right] &\leq \sum_{t=1}^k \mathbb{P}\left[z_3^t < v_t + z_2^t\right] \\
&\leq \sum_{t=1}^k \mathbb{P}\left[z_3^t < \frac{Qn}{k} - \frac{QN}{2k} + 1\right] + \sum_{t=1}^k \mathbb{P}\left[z_2^t > P\left(\frac{N}{k} - 1\right) + \frac{QN}{2k}\right] \\
&\leq k \cdot \exp\left(-\frac{(QN/2k-1)^2}{3Qn/k}\right) + k \cdot \exp\left(-\frac{(QN/2k)^2}{2P(N/k-1)}\right) \\
&\leq 2k \cdot \exp\left(-\frac{Q^2N^2}{48Pkn}\right)
\end{aligned}$$

where the third inequality follows from standard Chernoff bounds on the tails of the binomial distribution. Marginalizing this bound over $v \sim \mathcal{L}(z_1) = \text{Bern}(P)^{\otimes k}$, we have that

$$\mathbb{P}\left[\mathcal{E}^C\right] = \mathbb{E}_{v \sim \mathcal{L}(z_1)} \mathbb{P}\left[\mathcal{E}^C \middle| z_1 = v\right] \leq 2k \cdot \exp\left(-\frac{Q^2N^2}{48Pkn}\right)$$

Now consider the total variation error induced by conditioning each of the product measures $\otimes_{t=1}^k \mathcal{L}(z_1^t + z_2^t + z_4^t)$ and $\otimes_{t=1}^k \mathcal{L}(z_3^t)$ on the event $\mathcal{E}$. Note that under $\mathcal{E}$, by definition, we have that $z_3^t = z_1^t + z_2^t + z_4^t$ for each $t \in [k]$. By the conditioning property of d_{TV} in Fact 15, we have

$$d_{\text{TV}}\left(\otimes_{t=1}^k \mathcal{L}(z_1^t + z_2^t + z_4^t), \mathcal{L}\left((z_3^t : t \in [k]) \middle| \mathcal{E}\right)\right) \leq \mathbb{P}\left[\mathcal{E}^C\right]$$
$$d_{\text{TV}}\left(\otimes_{t=1}^k \mathcal{L}(z_3^t), \mathcal{L}\left((z_3^t : t \in [k]) \middle| \mathcal{E}\right)\right) \leq \mathbb{P}\left[\mathcal{E}^C\right]$$

The fact that $\otimes_{t=1}^k \mathcal{L}(z_3^t) = \text{Bin}(n/k, Q)^{\otimes k}$ and the triangle inequality now imply that

$$d_{\text{TV}}\left(\otimes_{t=1}^k \mathcal{L}(z_1^t + z_2^t + z_4^t), \text{Bin}(n/k, Q)^{\otimes k}\right) \leq 2 \cdot \mathbb{P}\left[\mathcal{E}^C\right] \leq 4k \cdot \exp\left(-\frac{Q^2N^2}{48Pkn}\right)$$

which proves the second inequality in the statement of the lemma. It suffices to establish the first inequality. A similar conditioning step as above shows that for all $v \in \{0,1\}^k$, we have that

$$d_{\mathrm{TV}}\left(\otimes_{t=1}^k \mathcal{L}\left(v_t + z_2^t + z_4^t \middle| z_1^t = v_t\right), \mathcal{L}\left(\left(v_t + z_2^t + z_4^t : t \in [k]\right) \middle| z_1 = v \text{ and } \mathcal{E}\right)\right) \leq \mathbb{P}\left[\mathcal{E}^C \middle| z_1 = v\right]$$
$$d_{\mathrm{TV}}\left(\otimes_{t=1}^k \mathcal{L}\left(z_3^t \middle| z_1^t = v_t\right), \mathcal{L}\left(\left(z_3^t : t \in [k]\right) \middle| z_1 = v \text{ and } \mathcal{E}\right)\right) \leq \mathbb{P}\left[\mathcal{E}^C \middle| z_1 = v\right]$$

The triangle inequality and the fact that $z_3 \sim \mathrm{Bin}(n/k, Q)^{\otimes k}$ is independent of z_1 implies that

$$d_{\mathrm{TV}}\left(\otimes_{t=1}^k \mathcal{L}\left(v_t + z_2^t + z_4^t \middle| z_1^t = v_t\right), \mathrm{Bin}(n/k, Q)^{\otimes k}\right) \leq 4k \cdot \exp\left(-\frac{Q^2N^2}{48Pkn}\right)$$

By Lemma 17 applied with $P_t = v_t \in \{0,1\}$, we also have that

$$d_{\mathrm{TV}}\left(\otimes_{t=1}^k \left(v_t + \mathrm{Bin}(n/k-1, Q)\right), \mathrm{Bin}(n/k, Q)^{\otimes k}\right) \leq \sqrt{\sum_{t=1}^k \frac{k(v_t - Q)^2}{2nQ(1-Q)}} \leq \sqrt{\frac{C_Q k^2}{2n}}$$

The triangle now implies that for each $v \in \{0,1\}^k$,

$$\begin{aligned}
&d_{\mathrm{TV}}\left(\otimes_{t=1}^k \mathcal{L}\left(z_2^t + z_4^t \middle| z_1^t = v_t\right), \mathrm{Bin}(n/k-1, Q)^{\otimes k}\right) \\
&\quad = d_{\mathrm{TV}}\left(\otimes_{t=1}^k \mathcal{L}\left(v_t + z_2^t + z_4^t \middle| z_1^t = v_t\right), \otimes_{t=1}^k \left(v_t + \mathrm{Bin}(n/k-1, Q)\right)\right) \\
&\quad \leq 4k \cdot \exp\left(-\frac{Q^2N^2}{48Pkn}\right) + \sqrt{\frac{C_Q k^2}{2n}}
\end{aligned}$$

We now marginalize over $v \sim \mathcal{L}(z_1) = \mathrm{Bern}(P)^{\otimes k}$. The conditioning on a random variable property of d_{TV} in Fact 15 implies that

$$\begin{aligned}
&d_{\mathrm{TV}}\left(\otimes_{t=1}^k \mathcal{L}(z_1^t, z_2^t + z_4^t), \left(\mathrm{Bern}(P) \otimes \mathrm{Bin}(n/k-1, Q)\right)^{\otimes k}\right) \\
&\quad \leq \mathbb{E}_{v \sim \mathrm{Bern}(P)^{\otimes k}} d_{\mathrm{TV}}\left(\otimes_{t=1}^k \mathcal{L}\left(z_2^t + z_4^t \middle| z_1^t = v_t\right), \mathrm{Bin}(n/k-1, Q)^{\otimes k}\right)
\end{aligned}$$

which, when combined with the inequalities above, completes the proof of the lemma. ■

We now apply this lemma to prove Lemma 23. The proof of this lemma is a k-partite variant of the argument used to prove Theorem 6.1 in Brennan et al. (2019a). However, it involves several technical subtleties that do not arise in the non k-partite case.

Proof (of Lemma 23) Fix some subset $R \subseteq [N]$ such that $|R \cap E_i| = 1$ for each $i \in [k]$. We will first show that $\mathcal{A}$ maps an input $G \sim \mathcal{G}(N, R, p, q)$ approximately in total variation to a sample from the planted submatrix distribution $\mathcal{M}_{[n]\times[n]}\left(\mathcal{U}_n(F), \mathrm{Bern}(p), \mathrm{Bern}(Q)\right)$. By AM-GM, we have that

$$\sqrt{pq} \leq \frac{p+q}{2} = 1 - \frac{(1-p)+(1-q)}{2} \leq 1 - \sqrt{(1-p)(1-q)}$$

If $p \neq 1$, it follows that $P = p > Q = 1 - \sqrt{(1-p)(1-q)}$. This implies that $\frac{1-p}{1-q} = \left(\frac{1-P}{1-Q}\right)^2$ and the inequality above rearranges to $\left(\frac{P}{Q}\right)^2 \leq \frac{p}{q}$. If $p = 1$, then $Q = \sqrt{q}$ and $\left(\frac{P}{Q}\right)^2 = \frac{p}{q}$. Furthermore,

the inequality $\frac{1-p}{1-q} \leq \left(\frac{1-P}{1-Q}\right)^2$ holds trivially. Therefore we may apply Lemma 21, which implies that $(G_1, G_2) \sim \mathcal{G}(N, R, p, Q)^{\otimes 2}$.

Let the random set $U = \{\pi_1^{-1}(R \cap E_1), \pi_2^{-1}(R \cap E_2), \ldots, \pi_k^{-1}(R \cap E_k)\}$ denote the support of the k-subset of $[n]$ that R is mapped to in the embedding step of TO-k-PARTITE-SUBMATRIX. Now fix some k-subset $R' \subseteq [n]$ with $|R' \cap F_i| = 1$ for each $i \in [k]$ and consider the distribution of M_{PD} conditioned on the event $U = R'$. Since $(G_1, G_2) \sim \mathcal{G}(n, R, p, Q)^{\otimes 2}$, Step 2 of TO-$k$-PARTITE-SUBMATRIX ensures that the off-diagonal entries of M_{PD}, given this conditioning, are independent and distributed as follows:

- $M_{ij} \sim \text{Bern}(p)$ if $i \neq j$ and $i, j \in R'$; and
- $M_{ij} \sim \text{Bern}(Q)$ if $i \neq j$ and $i \notin R'$ or $j \notin R'$.

which match the corresponding entries of $\mathcal{M}_{[n]\times[n]}(R' \times R', \text{Bern}(p), \text{Bern}(Q))$. Furthermore, these entries are independent of the vector $\text{diag}(M_{\text{PD}}) = ((M_{\text{PD}})_{ii} : i \in [k])$ of the diagonal entries of M_{PD}. It therefore follows that

$$\begin{aligned} &d_{\text{TV}}\left(\mathcal{L}\left(M_{\text{PD}} \middle| U = R'\right), \mathcal{M}_{[n]\times[n]}\left(R' \times R', \text{Bern}(p), \text{Bern}(Q)\right)\right) \\ &\quad = d_{\text{TV}}\left(\mathcal{L}\left(\text{diag}(M_{\text{PD}}) \middle| U = R'\right), \mathcal{M}_{[n]}\left(R', \text{Bern}(p), \text{Bern}(Q)\right)\right) \end{aligned}$$

Let $(S_1', S_2', \ldots, S_k')$ be any tuple of fixed subsets such that $|S_t'| = N/k$, $S_i' \subseteq F_t$ and $R' \cap F_t \in S_t'$ for each $t \in [k]$. Now consider the distribution of $\text{diag}(M_{\text{PD}})$ conditioned on both $U = R'$ and $(S_1, S_2, \ldots, S_k) = (S_1', S_2', \ldots, S_k')$. It holds by construction that the k vectors $\text{diag}(M_{\text{PD}})_{F_t}$ are independent for $t \in [k]$ and each distributed as follows:

- $\text{diag}(M_{\text{PD}})_{S_t'}$ is an exchangeable distribution on $\{0,1\}^{N/k}$ with support of size $s_1^t \sim \text{Bin}(N/k, p)$, by construction. This implies that $\text{diag}(M_{\text{PD}})_{S_t'} \sim \text{Bern}(p)^{\otimes N/k}$. This can trivially be restated as $\left(M_{R'\cap F_t, R'\cap F_t}, \text{diag}(M_{\text{PD}})_{S_t'\setminus R'}\right) \sim \text{Bern}(p) \otimes \text{Bern}(p)^{\otimes N/k-1}$.
- $\text{diag}(M_{\text{PD}})_{F_t\setminus S_t'}$ is an exchangeable distribution on $\{0,1\}^{N/k}$ with support of size $z_4^t = \max\{s_2^t - s_1^t, 0\}$. Furthermore, $\text{diag}(M_{\text{PD}})_{F_t\setminus S_t'}$ is independent of $\text{diag}(M_{\text{PD}})_{S_t'}$.

For each $t \in [k]$, let $z_1^t = M_{R'\cap F_t, R'\cap F_t} \sim \text{Bern}(p)$ and $z_2^t \sim \text{Bin}(N/k - 1, p)$ be the size of the support of $\text{diag}(M_{\text{PD}})_{S_t'\setminus R'}$. As shown discussed in the first point above, we have that z_1^t and z_2^t are independent and $z_1^t + z_2^t = s_1^t$.

Now consider the distribution of $\text{diag}(M_{\text{PD}})$ relaxed to only be conditioned on $U = R'$, and no longer on $(S_1, S_2, \ldots, S_k) = (S_1', S_2', \ldots, S_k')$. Conditioned on $U = R'$, the S_t are independent and each uniformly distributed among all N/k size subsets of F_t that contain the element $R' \cap F_t$. In particular, this implies that the distribution of $\text{diag}(M_{\text{PD}})_{F_t\setminus R'}$ is an exchangeable distribution on $\{0,1\}^{n/k-1}$ with support size $z_2^t + z_4^t$ for each t. Note that any $v \sim \mathcal{M}_{[n]}(R', \text{Bern}(p), \text{Bern}(Q))$ also satisfies that $v_{F_t\setminus R'}$ is exchangeable. This implies that $\mathcal{M}_{[n]}(R', \text{Bern}(p), \text{Bern}(Q))$ and $\text{diag}(M_{\text{PD}})$ are identically distributed when conditioned on their entries with indices in R' and on their support

sizes within the k sets of indices $F_t \backslash R'$. The conditioning property of Fact 15 therefore implies that

$$\begin{aligned} &d_{\mathrm{TV}}\left(\mathcal{L}\left(\mathrm{diag}(M_{\mathrm{PD}})\middle| U = R'\right), \mathcal{M}_{[n]}\left(R', \mathrm{Bern}(p), \mathrm{Bern}(Q)\right)\right) \\ &\quad \le d_{\mathrm{TV}}\left(\otimes_{t=1}^{k} \mathcal{L}(z_1^t, z_2^t + z_4^t), (\mathrm{Bern}(p) \otimes \mathrm{Bin}(n/k - 1, Q))^{\otimes k}\right) \\ &\quad \le 4k \cdot \exp\left(-\frac{Q^2 N^2}{48 P k n}\right) + \sqrt{\frac{C_Q k^2}{2n}} \end{aligned}$$

by the first inequality in Lemma 108. Now observe that $U \sim \mathcal{U}_n(F)$ and thus marginalizing over $R' \sim \mathcal{L}(U) = \mathcal{U}_n(F)$ and applying the conditioning property of Fact 15 yields that

$$\begin{aligned} &d_{\mathrm{TV}}\left(\mathcal{A}(G(N, R, p, q)), \mathcal{M}_{[n]\times[n]}\left(\mathcal{U}_n(F), \mathrm{Bern}(p), \mathrm{Bern}(Q)\right)\right) \\ &\quad \le \mathbb{E}_{R' \sim \mathcal{U}_n(F)}\, d_{\mathrm{TV}}\left(\mathcal{L}\left(M_{\mathrm{PD}}\middle| U = R'\right), \mathcal{M}_{[n]\times[n]}\left(R' \times R', \mathrm{Bern}(p), \mathrm{Bern}(Q)\right)\right) \end{aligned}$$

since $M_{\mathrm{PD}} \sim \mathcal{A}(\mathcal{G}(N, R, p, q))$. Applying an identical marginalization over $R \sim \mathcal{U}_N(E)$ completes the proof of the first inequality in the lemma statement.

It suffices to consider the case where $G \sim \mathcal{G}(N, q)$, which follows from an analogous but simpler argument. By Lemma 21, we have that $(G_1, G_2) \sim \mathcal{G}(N, Q)^{\otimes 2}$. It follows that the entries of M_{PD} are distributed as $(M_{\mathrm{PD}})_{ij} \sim_{\text{i.i.d.}} \mathrm{Bern}(Q)$ for all $i \neq j$ independently of $\mathrm{diag}(M_{\mathrm{PD}})$. Now note that the k vectors $\mathrm{diag}(M_{\mathrm{PD}})_{F_t}$ for $t \in [k]$ are each exchangeable and have support size $s_1^t + \max\{s_2^t - s_1^t, 0\} = z_1^t + z_2^t + z_4^t$ where $z_1^t \sim \mathrm{Bern}(p)$, $z_2^t \sim \mathrm{Bin}(N/k - 1, p)$ and $s_2^t \sim \mathrm{Bin}(n/k, Q)$ are independent. By the same argument as above, we have that

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathcal{L}(M_{\mathrm{PD}}), \mathrm{Bern}(Q)^{\otimes n\times n}\right) &= d_{\mathrm{TV}}\left(\mathcal{L}(\mathrm{diag}(M_{\mathrm{PD}})), \mathrm{Bern}(Q)^{\otimes n}\right) \\ &= d_{\mathrm{TV}}\left(\otimes_{t=1}^{k} \mathcal{L}\left(z_1^t + z_2^t + z_4^t\right), \mathrm{Bin}(n/k, Q)\right) \\ &\le 4k \cdot \exp\left(-\frac{Q^2 N^2}{48 P k n}\right) \end{aligned}$$

by Lemma 108. Since $M_{\mathrm{PD}} \sim \mathcal{A}(\mathcal{G}(N, q))$, this completes the proof of the lemma. ∎

Q.3. Proofs for Symmetric 3-ary Rejection Kernels

In this section, we establish the approximate Markov transition properties for symmetric 3-ary rejection kernels introduced in Section F.3.

Proof (of Lemma 25) Define $\mathcal{L}_1, \mathcal{L}_2 : X \to \mathbb{R}$ to be

$$\mathcal{L}_1(x) = \frac{d\mathcal{P}_+}{d\mathcal{Q}}(x) - \frac{d\mathcal{P}_-}{d\mathcal{Q}}(x) \quad \text{and} \quad \mathcal{L}_2(x) = \frac{d\mathcal{P}_+}{d\mathcal{Q}}(x) + \frac{d\mathcal{P}_-}{d\mathcal{Q}}(x) - 2$$

Note that if $x \in S$, then the triangle inequality implies that

$$\begin{aligned} P_A(x, 1) &\le \frac{1}{2}\left(1 + \frac{a}{4|\mu_2|} \cdot |\mathcal{L}_2(x)| + \frac{1}{4|\mu_1|} \cdot |\mathcal{L}_1(x)|\right) \le 1 \\ P_A(x, 1) &\ge \frac{1}{2}\left(1 - \frac{a}{4|\mu_2|} \cdot |\mathcal{L}_2(x)| - \frac{1}{4|\mu_1|} \cdot |\mathcal{L}_1(x)|\right) \ge 0 \end{aligned}$$

Similar computations show that $0 \le P_A(x,0) \le 1$ and $0 \le P_A(x,-1) \le 1$, implying that each of these probabilities is well-defined. Now let $R_1 = \mathbb{P}_{X\sim\mathcal{P}_+}[X \in S]$, $R_0 = \mathbb{P}_{X\sim\mathcal{Q}}[X \in S]$ and $R_{-1} = \mathbb{P}_{X\sim\mathcal{P}_-}[X \in S]$ where $R_1, R_0, R_{-1} \ge 1-\delta$ by assumption.

We now define several useful events. For the sake of analysis, consider continuing to iterate Step 2 even after z is set for the first time for a total of N iterations. Let A_i^1, A_i^0 and A_i^{-1} be the events that z is set in the ith iteration of Step 2 when $B=1$, $B=0$ and $B=-1$, respectively. Let $B_i^1 = (A_1^1)^C \cap (A_2^1)^C \cap \cdots \cap (A_{i-1}^1)^C \cap A_i^1$ be the event that z is set for the first time in the ith iteration of Step 2. Let $C^1 = A_1^1 \cup A_2^1 \cup \cdots \cup A_N^1$ be the event that z is set in some iteration of Step 2. Define B_i^0, C^0, B_i^{-1} and C^{-1} analogously. Let z_0 be the initialization of z in Step 1.

Now let $Z_1 \sim \mathcal{D}_1 = \mathcal{L}(3\text{-SRK}(1))$, $Z_0 \sim \mathcal{D}_0 = \mathcal{L}(3\text{-SRK}(0))$ and $Z_{-1} \sim \mathcal{D}_{-1} = \mathcal{L}(3\text{-SRK}(-1))$. Note that $\mathcal{L}(Z_t|B_i^t) = \mathcal{L}(Z_t|A_i^t)$ for each $t \in \{-1,0,1\}$ since A_i^t is independent of $A_1^t, A_2^t, \ldots, A_{i-1}^t$ and the sample z' chosen in the ith iteration of Step 2. The independence between Steps 2.1 and 2.3 implies that

$$\begin{aligned}
\mathbb{P}\left[A_i^1\right] &= \mathbb{E}_{x\sim\mathcal{Q}}\left[\frac{1}{2}\left(1 + \frac{a}{4\mu_2}\cdot\mathcal{L}_2(x) + \frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\cdot\mathbf{1}_S(x)\right] \\
&= \frac{1}{2}R_0 + \frac{a}{8\mu_2}\left(R_1 + R_{-1} - 2R_0\right) + \frac{1}{8\mu_1}\left(R_1 - R_{-1}\right) \ge \frac{1}{2} - \frac{\delta}{2}\left(1 + \frac{a}{2}|\mu_2|^{-1} + \frac{1}{4}|\mu_1|^{-1}\right) \\
\mathbb{P}\left[A_i^0\right] &= \mathbb{E}_{x\sim\mathcal{Q}}\left[\frac{1}{2}\left(1 - \frac{1-a}{4\mu_2}\cdot\mathcal{L}_2(x)\right)\cdot\mathbf{1}_S(x)\right] \\
&= \frac{1}{2}R_0 - \frac{1-a}{8\mu_2}\left(R_1 + R_{-1} - 2R_0\right) \ge \frac{1}{2} - \frac{\delta}{2}\left(1 + \frac{1-a}{4}\cdot|\mu_2|^{-1}\right) \\
\mathbb{P}\left[A_i^{-1}\right] &= \mathbb{E}_{x\sim\mathcal{Q}}\left[\frac{1}{2}\left(1 + \frac{a}{4\mu_2}\cdot\mathcal{L}_2(x) - \frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\cdot\mathbf{1}_S(x)\right] \\
&= \frac{1}{2}R_0 + \frac{a}{8\mu_2}\left(R_1 + R_{-1} - 2R_0\right) - \frac{1}{4\mu_1}\left(R_1 - R_{-1}\right) \ge \frac{1}{2} - \frac{\delta}{2}\left(1 + \frac{a}{2}|\mu_2|^{-1} + \frac{1}{4}|\mu_1|^{-1}\right)
\end{aligned}$$

The independence of the A_i^t for each $t \in \{-1,0,1\}$ implies that

$$1 - \mathbb{P}\left[C^t\right] = \prod_{i=1}^N\left(1 - \mathbb{P}\left[A_i^t\right]\right) \le \left(\frac{1}{2} + \frac{\delta}{2}\left(1 + \frac{1}{2}|\mu_2|^{-1} + |\mu_1|^{-1}\right)\right)^N$$

Note that $\mathcal{L}(Z_t|A_i^t)$ are each absolutely continuous with respect to $\mathcal{Q}$ or each $t \in \{-1,0,1\}$, with Radon-Nikodym derivatives given by

$$\begin{aligned}
\frac{d\mathcal{L}(Z_1|B_i^1)}{d\mathcal{Q}}(x) &= \frac{d\mathcal{L}(Z_1|A_i^1)}{d\mathcal{Q}}(x) = \frac{1}{2\cdot\mathbb{P}\left[A_i^1\right]}\left(1 + \frac{a}{4\mu_2}\cdot\mathcal{L}_2(x) + \frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\cdot\mathbf{1}_S(x) \\
\frac{d\mathcal{L}(Z_0|B_i^0)}{d\mathcal{Q}}(x) &= \frac{d\mathcal{L}(Z_0|A_i^0)}{d\mathcal{Q}}(x) = \frac{1}{2\cdot\mathbb{P}\left[A_i^1\right]}\left(1 - \frac{1-a}{4\mu_2}\cdot\mathcal{L}_2(x)\right)\cdot\mathbf{1}_S(x) \\
\frac{d\mathcal{L}(Z_{-1}|B_i^{-1})}{d\mathcal{Q}}(x) &= \frac{d\mathcal{L}(Z_{-1}|A_i^{-1})}{d\mathcal{Q}}(x) = \frac{1}{2\cdot\mathbb{P}\left[A_i^1\right]}\left(1 + \frac{a}{4\mu_2}\cdot\mathcal{L}_2(x) - \frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\cdot\mathbf{1}_S(x)
\end{aligned}$$

Fix one of $t \in \{-1,0,1\}$ and note that since the conditional laws $\mathcal{L}(Z_t|B_i^t)$ are all identical, we have that

$$\frac{d\mathcal{D}_t}{d\mathcal{Q}}(x) = \mathbb{P}\left[C^t\right]\cdot\frac{d\mathcal{L}(Z_t|B_1^t)}{d\mathcal{Q}}(x) + \left(1 - \mathbb{P}\left[C^t\right]\right)\cdot\mathbf{1}_{z_0}(x)$$

Therefore it follows that

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathcal{D}_t, \mathcal{L}(Z_t|B_1^t)\right) &= \frac{1}{2}\cdot\mathbb{E}_{x\sim\mathcal{Q}}\left[\left|\frac{d\mathcal{D}_t}{d\mathcal{Q}}(x) - \frac{d\mathcal{L}(Z_t|B_1^t)}{d\mathcal{Q}}(x)\right|\right] \\ &\le \frac{1}{2}\left(1-\mathbb{P}\left[C^t\right]\right)\cdot\mathbb{E}_{x\sim\mathcal{Q}}\left[\mathbf{1}_{z_0}(x) + \frac{d\mathcal{L}(Z_t|B_1^t)}{d\mathcal{Q}}(x)\right] = 1-\mathbb{P}\left[C^t\right] \end{aligned}$$

by the triangle inequality. Since $1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)+\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\ge 0$ for $x\in S$, we have that

$$\begin{aligned} &\mathbb{E}_{x\sim\mathcal{Q}}\left[\left|\frac{d\mathcal{L}(Z_1|B_1^1)}{d\mathcal{Q}}(x) - \left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)+\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\right|\right] \\ &\quad= \left|\frac{1}{2\cdot\mathbb{P}\left[A_i^1\right]}-1\right|\cdot\mathbb{E}_{x\sim\mathcal{Q}_n^*}\left[\left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)+\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\cdot\mathbf{1}_S(x)\right] \\ &\qquad+\mathbb{E}_{x\sim\mathcal{Q}}\left[\left|1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)+\frac{1}{4|\mu_1|}\cdot\mathcal{L}_1(x)\right|\cdot\mathbf{1}_{S^C}(x)\right] \\ &\quad\le \left|\frac{1}{2}-\mathbb{P}[A_i^1]\right| + \mathbb{E}_{x\sim\mathcal{Q}}\left[\left(1+\frac{a}{4|\mu_2|}\cdot\left(\frac{d\mathcal{P}_+}{d\mathcal{Q}}(x)+\frac{d\mathcal{P}_-}{d\mathcal{Q}}(x)+2\right)\right)\cdot\mathbf{1}_{S^C}(x)\right] \\ &\qquad+\mathbb{E}_{x\sim\mathcal{Q}}\left[\frac{1}{4|\mu_1|}\cdot\left(\frac{d\mathcal{P}_+}{d\mathcal{Q}}(x)+\frac{d\mathcal{P}_-}{d\mathcal{Q}}(x)\right)\cdot\mathbf{1}_{S^C}(x)\right] \\ &\quad\le \frac{\delta}{2}\left(1+\frac{a}{2}|\mu_2|^{-1}+\frac{1}{4}|\mu_1|^{-1}\right)+\delta\left(1+a|\mu_2|^{-1}+\frac{1}{2}|\mu_1|^{-1}\right) = \delta\left(\frac{3}{2}+\frac{5}{4}|\mu_2|^{-1}+\frac{5}{8}|\mu_1|^{-1}\right) \end{aligned}$$

By analogous computations, we have that

$$\begin{aligned} \mathbb{E}_{x\sim\mathcal{Q}}\left[\left|\frac{d\mathcal{L}(Z_0|B_1^0)}{d\mathcal{Q}}(x) - \left(1-\frac{1-a}{4\mu_2}\cdot\mathcal{L}_2(x)\right)\right|\right] &\le 2\delta\left(1+|\mu_1|^{-1}+|\mu_2|^{-1}\right) \\ \mathbb{E}_{x\sim\mathcal{Q}}\left[\left|\frac{d\mathcal{L}(Z_{-1}|B_1^{-1})}{d\mathcal{Q}}(x) - \left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)-\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\right|\right] &\le 2\delta\left(1+|\mu_1|^{-1}+|\mu_2|^{-1}\right) \end{aligned}$$

Now observe that

$$\begin{aligned} \frac{d\mathcal{P}_+}{d\mathcal{Q}}(x) &= \left(\frac{1-a}{2}+\mu_1+\mu_2\right)\cdot\left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)+\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)+(a-2\mu_2)\cdot\left(1-\frac{1-a}{4\mu_2}\cdot\mathcal{L}_2(x)\right) \\ &\quad+\left(\frac{1-a}{2}-\mu_1+\mu_2\right)\cdot\left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)-\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right) \\ 1 &= \frac{1-a}{2}\cdot\left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)+\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)+a\cdot\left(1-\frac{1-a}{4\mu_2}\cdot\mathcal{L}_2(x)\right) \\ &\quad+\frac{1-a}{2}\cdot\left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)-\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right) \\ \frac{d\mathcal{P}_-}{d\mathcal{Q}}(x) &= \left(\frac{1-a}{2}-\mu_1+\mu_2\right)\cdot\left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)+\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)+(a-2\mu_2)\cdot\left(1-\frac{1-a}{4\mu_2}\cdot\mathcal{L}_2(x)\right) \\ &\quad+\left(\frac{1-a}{2}+\mu_1+\mu_2\right)\cdot\left(1+\frac{a}{4\mu_2}\cdot\mathcal{L}_2(x)-\frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right) \end{aligned}$$

Let $\mathcal{D}^*$ be the mixture of $\mathcal{L}(Z_1|B_1^1), \mathcal{L}(Z_0|B_1^0)$ and $\mathcal{L}(Z_{-1}|B_1^{-1})$ with weights $\frac{1-a}{2} + \mu_1 + \mu_2, a - 2\mu_2$ and $\frac{1-a}{2} - \mu_1 + \mu_2$, respectively. It then follows by the triangle inequality that

$$\begin{aligned}
&d_{\mathrm{TV}}\left(3\text{-SRK}(\mathrm{Tern}(a,\mu_1,\mu_2)), \mathcal{P}_+\right) \\
&\quad\le d_{\mathrm{TV}}\left(\mathcal{D}^*, \mathcal{P}_+\right) + d_{\mathrm{TV}}\left(\mathcal{D}^*, 3\text{-SRK}(\mathrm{Tern}(a,\mu_1,\mu_2))\right) \\
&\quad\le \left(\frac{1-a}{2} + \mu_1 + \mu_2\right) \cdot \mathbb{E}_{x\sim\mathcal{Q}}\left[\left|\frac{d\mathcal{L}(Z_1|B_1^1)}{d\mathcal{Q}}(x) - \left(1 + \frac{a}{4\mu_2}\cdot\mathcal{L}_2(x) + \frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\right|\right] \\
&\qquad + (a-2\mu_2)\cdot\mathbb{E}_{x\sim\mathcal{Q}}\left[\left|\frac{d\mathcal{L}(Z_0|B_1^0)}{d\mathcal{Q}}(x) - \left(1 - \frac{1-a}{4\mu_2}\cdot\mathcal{L}_2(x)\right)\right|\right] \\
&\qquad + \left(\frac{1-a}{2} - \mu_1 + \mu_2\right) \cdot \mathbb{E}_{x\sim\mathcal{Q}}\left[\left|\frac{d\mathcal{L}(Z_{-1}|B_1^{-1})}{d\mathcal{Q}}(x) - \left(1 + \frac{a}{4\mu_2}\cdot\mathcal{L}_2(x) - \frac{1}{4\mu_1}\cdot\mathcal{L}_1(x)\right)\right|\right] \\
&\qquad + \left(\frac{1-a}{2} + \mu_1 + \mu_2\right)\cdot d_{\mathrm{TV}}\left(\mathcal{D}_1, \mathcal{L}(Z_1|B_1^1)\right) + (a-2\mu_2)\cdot d_{\mathrm{TV}}\left(\mathcal{D}_1, \mathcal{L}(Z_0|B_1^0)\right) \\
&\qquad + \left(\frac{1-a}{2} - \mu_1 + \mu_2\right)\cdot d_{\mathrm{TV}}\left(\mathcal{D}_{-1}, \mathcal{L}(Z_{-1}|B_1^{-1})\right) \\
&\quad\le 2\delta\left(1 + |\mu_1|^{-1} + |\mu_2|^{-1}\right) + \left(\frac{1}{2} + \delta\left(1 + |\mu_1|^{-1} + |\mu_2|^{-1}\right)\right)^N
\end{aligned}$$

A symmetric argument shows analogous upper bounds on both $d_{\mathrm{TV}}\left(3\text{-SRK}(\mathrm{Tern}(a,-\mu_1,\mu_2)), \mathcal{P}_-\right)$ and $d_{\mathrm{TV}}\left(3\text{-SRK}(\mathrm{Tern}(a,0,0)), \mathcal{Q}\right)$, completing the proof of the lemma. ■

Q.4. Proofs for Label Generation

In this section, we give the two deferred proofs from Section I.2.

Proof (of Lemma 54) This lemma follows from a similar argument to Lemma 53. As in Lemma 53, the given conditions on C, γ, μ' and N imply that

$$2\left(\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\right)^2 \le 1$$

and thus X' is well-defined almost surely. First observe that if $Z = \mu''\cdot u + G'$ where $G' \sim \mathcal{N}(0, I_d)$ then

$$X' = \frac{a\gamma\cdot y'}{1+\gamma^2}\cdot u + \frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\cdot G' + \frac{1}{\sqrt{2}}\cdot\sqrt{1 - 2\left(\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\right)^2}\cdot G + \frac{1}{\sqrt{2}}\cdot W$$

where $a = \mu''/\mu'$. Thus by the same argument as in Lemma 53, we have that

$$\mathcal{L}(X'|y') = \mathcal{N}\left(\frac{a\gamma\cdot y}{1+\gamma^2}\cdot u,\ I_d - \frac{\gamma^2}{1+\gamma^2}\cdot uu^\top\right)$$

Now note that by the conditioning property of multivariate Gaussians, we have that

$$\mathcal{L}(X|y) = \mathcal{N}\left(\Sigma_{Xy}\Sigma_{yy}^{-1}\cdot y,\ \Sigma_{XX} - \Sigma_{Xy}\Sigma_{yy}^{-1}\Sigma_{yX}\right)$$

It is easily verified that

$$\Sigma_{Xy}\Sigma_{yy}^{-1} = \frac{a\gamma}{1+\gamma^2}\cdot u \quad \text{and} \quad \Sigma_{XX} - \Sigma_{Xy}\Sigma_{yy}^{-1}\Sigma_{yX} = I_d - \frac{\gamma^2}{1+\gamma^2}\cdot uu^\top$$

and thus $\mathcal{L}(X|y)$ and $\mathcal{L}(X'|y')$ are equidistributed. Since $y \sim \mathcal{N}(0, 1+\gamma^2)$, it follows by the same application of the conditioning property in Fact 15 as in Lemma 53 implies that

$$d_{\mathrm{TV}}\left(\mathcal{L}(X,y), \mathcal{L}(X',y')\right) \le d_{\mathrm{TV}}\left(\mathcal{L}(y), \mathcal{L}(y')\right) = O\left(N^{-C^2/2}\right)$$

which completes the proof of the lemma. ■

Proof (of Lemma 55) This lemma follows from a similar argument to Lemma 53. As in Lemmas 53 and 54, the given conditions imply that X' is well-defined almost surely. Conditioned on y', it holds that Z, G and W are independent. Therefore the three terms in the definition of X' are independent and distributed as

$$\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\cdot Z \sim \mathcal{N}\left(0, \left(\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\right)^2\cdot I_d\right),$$
$$\frac{1}{\sqrt{2}}\cdot\sqrt{1-2\left(\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\right)^2}\cdot G \sim \mathcal{N}\left(0, \frac{1}{2}\cdot I_d - \left(\frac{\gamma\cdot y'}{\mu'(1+\gamma^2)}\right)^2\cdot I_d\right) \quad \text{and}$$
$$\frac{1}{\sqrt{2}}\cdot W \sim \mathcal{N}\left(0, \frac{1}{2}\cdot I_d\right)$$

conditioned on y'. It follows that $X'|y' \sim \mathcal{N}(0, I_d)$ and thus X' is independent of y'. Now let $X \in \mathbb{R}^d$ and $y \in \mathbb{R}$ be such that $X \sim \mathcal{N}(0, I_d)$ and $y \sim \mathcal{N}(0, 1+\gamma^2)$ are independent. The same application of the conditioning property in Fact 15 as in Lemmas 53 and 54 now completes the proof of the lemma. ■

Appendix R. Deferred Proofs from Part III

R.1. Proofs from Secret Leakage and the PC_ρ Conjecture

In this section, we present the deferred proof of Lemma 75 from Section K. The proof of this lemma is similar to the proof of Lemma 5.2 in Feldman et al. (2013).

Proof (of Lemma 75) The proof is almost identical to Lemma 5.2 in Feldman et al. (2013) and we give a sketch here. Lemma 74 implies that $\sum_{T\in A}\left|\langle \widehat{D}_S, \widehat{D}_T\rangle_D\right| \le \sum_{T\in A} 2^{|S\cap T|}k^2/n^2$. If the only constraint on A is its cardinality, then the maximum value for the RHS is obtained by adding S to A, next $\{T : |T\cap S| = k-1\}$, and so forth with decreasing size of $|T\cap S|$, and we assume that A is defined in this manner. Letting $T_\lambda = \{T : |T\cap S| = \lambda\}|$, set $\lambda_0 = \min\{\lambda : T_\lambda \neq \varnothing\}$ so that $T_\lambda \subseteq A$ for $\lambda > \lambda_0$. We bound the ratio

$$\frac{|T_j|}{|T_{j+1}|} = \frac{\binom{k}{j}\left(\frac{n}{k}\right)^{k-j}}{\binom{k}{j+1}\left(\frac{n}{k}\right)^{k-j-1}} \ge \frac{jn}{k^2} = jn^{2\delta} \quad \text{hence} \quad |T_j| \le \frac{|T_0|}{(j-1)!n^{2\delta j}} \le \frac{|\mathcal{S}|}{(j-1)!n^{2\delta j}}\,.$$

Now

$$|A| \leq \sum_{j \geq \lambda_0} |T_j| \leq |\mathcal{S}| n^{-2\delta\lambda_0} \sum_{j \geq \lambda_0} \frac{1}{(j-1)! n^{2\delta(j-\lambda_0)}} \leq 2|\mathcal{S}| n^{-2\delta\lambda_0}$$

for n greater than some constant. Thus if $|A| \geq 2|\mathcal{S}|/n^{2\ell\delta}$, we must conclude that $\ell \geq \lambda_0$. We bound the quantity $\sum_{T \in A} 2^{|S \cap T|} \leq \sum_{j=\lambda_0}^{k} 2^j |T_j \cap A| \leq 2^{\lambda_0} |T_{\lambda_0} \cap A| + \sum_{j=\lambda_0+1}^{k} 2^j |T_j| \leq 2^{\lambda_0}|A| + 2^{\lambda_0+2}|T_{\lambda_0+1}| \leq 2^{\lambda_0+3}|A| \leq 2^{\ell+3}|A|$. Here we used that $|T_{j+1}| \leq |T_j| n^{-2\delta}$ to bound by a geometric series and also that $T_{\lambda_0+1} \subseteq A$. Rearranging and combining with the inequality at the start of the proof concludes the argument. ■

R.2. Proofs for Reductions and Computational Lower Bounds

In this section, we present a number of deferred proofs from Part III. The majority of these proofs are similar to other proofs presented in the main body of the paper.

Proof (of Theorem 9) To prove this theorem, we will to show that Theorem 52 implies that k-BPDS-TO-MSLR applied with $r > 2$ fills out all of the possible growth rates specified by the computational lower bound $n = \tilde{o}(k^2\epsilon^2/\tau^4)$ and the other conditions in the theorem statement. As discussed above, it suffices to reduce in total variation to $\text{MSLR}(n, k, d, \tau, 1/r)$ where $1/r \leq \epsilon$.

Fix a constant pair of probabilities $0 < q < p \leq 1$ and any sequence of parameters $(n, k, d, \tau, \epsilon)$ all of which are implicitly functions of n such that (n, ϵ^{-1}) satisfies (T) and $(n, k, d, \tau, \epsilon)$ satisfy the conditions

$$n \leq c \cdot \frac{k^2\epsilon^2}{w^2 \cdot \tau^4 \cdot (\log n)^{4+2c'}}, \quad wk \leq n^{1/6} \quad \text{and} \quad wk^2 \leq d$$

for sufficiently large n, an arbitrarily slow-growing function $w = w(n) \to \infty$ at least satisfying that $w(n) = n^{o(1)}$, a sufficiently small constant $c > 0$ and a sufficiently large constant $c' > 0$. The rest of this proof will follow that of Theorem 4 very closely. In order to fulfill the criteria in Condition E.1, we specify M, N, k_M, k_N and n' exactly as in Theorem 4. As in Theorem 4, we have the inequalities

$$n' \leq w^{-2} r^{2t} = O\left(\frac{r^{2t}}{n} \cdot \frac{k^2\epsilon^2}{\tau^4 \cdot (\log n)^{2+2c'}}\right)$$

$$\tau \leq \frac{c^{1/4}\epsilon^{1/2}k^{1/2}}{n^{1/4}(\log n)^{(2+c')/2}} = \Theta\left(\frac{r^{t/2}}{n^{1/4}} \cdot \frac{k_M^{1/2}}{\sqrt{r^{t+1}(\log n)^{2+c'}}}\right)$$

Furthermore, we also have that

$$\tau^2 \leq \frac{c^{1/2} \cdot k}{w n^{1/2} \cdot (\log n)^{2+c'}} = O\left(\frac{r^t}{n} \cdot \frac{k_N k_M}{N \log(MN)}\right)$$

As long as $\sqrt{n} = \tilde{\Theta}(r^t)$ then: (2.1) the inequality above on n' would imply that $(n', k, d, \tau, \epsilon)$ is in the desired hard regime; (2.2) n and n' have the same growth rate since $w = n^{o(1)}$; and (2.3) $n \gg M^3$, $d \geq M$ and taking c' large enough would imply that τ satisfies the bounds needed to apply Theorem 52 to yield the desired reduction. By Lemma 80, there is an infinite subsequence of the input parameters such that $\sqrt{n} = \tilde{\Theta}(r^t)$, which concludes the proof as in Theorem 4. ■

Proof (of Lemma 91) First suppose that $M \sim \text{GHPM}_D(n, r, C, D, \gamma)$ where C and D are each sequences of r disjoint sets of size K. Since the M_{ij} are independent for $1 \le i, j \le n$, we now have that

$$\mathbb{E}[s_C(M)] = \sum_{i,j=1}^{n} \mathbb{E}\left[M_{ij}^2 - 1\right] = rK^2 \cdot \gamma^2 + \frac{rK^2}{r-1} \cdot \gamma^2$$

$$\text{Var}\left[s_C(M)\right] = \sum_{i,j=1}^{n} \text{Var}\left[M_{ij}^2 - 1\right] = rK^2 \cdot 4\gamma^2 + \frac{rK^2}{(r-1)^3} \cdot \gamma^2 + 2n^2$$

Here, we have used the following facts. If $X \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}[(\gamma + X)^2 - 1] = \gamma^2, \quad \mathbb{E}\left[\left(\frac{\gamma}{r-1} + X\right)^2 - 1\right] = \frac{\gamma^2}{(r-1)^2}$$

$$\text{Var}[X^2 - 1] = 2, \quad \text{Var}[(\gamma + X)^2 - 1] = 4\gamma^2 + 2, \quad \text{Var}\left[\left(\frac{\gamma}{r-1} + X\right)^2 - 1\right] = \frac{\gamma^2}{(r-1)^4} + 2$$

Note that $s_C(M)$ is invariant to permuting the rows and columns of M and thus $s_C(M)$ is equidistirbuted under $M \sim \text{GHPM}_D(n, r, C, D, \gamma)$ and $M \sim \text{GHPM}_D(n, r, K, \gamma)$. Now Chebyshev's inequality implies the desired lower bound on $s_C(M)$ in (1) holds with probability $1 - o_n(1)$. Now observe that

$$s_I(M) \ge \sum_{h=1}^{r} \sum_{i \in C_h} \sum_{j \in D_h} M_{ij} = Y$$

holds almost surely by definition when $M \sim \text{GHPM}_D(n, r, C, D, \gamma)$. Note that $Y \sim \mathcal{N}(rK^2\gamma, rK^2)$ conditioned on C and D and therefore it holds that $Y \ge rK^2\gamma - wr^{1/2}K$ with probability $1 - o_n(1)$. The second lower bound in (1) now follows since $s_I(M)$ is equidistirbuted under $M \sim \text{GHPM}_D(n, r, C, D, \gamma)$ and $M \sim \text{GHPM}_D(n, r, K, \gamma)$.

Now suppose that $M \sim \mathcal{N}(0,1)^{\otimes n \times n}$. In this case, $s_C(M) + n^2$ is distributed as $\chi^2(n^2)$ and the first upper bound in (2) holds by Chebyshev's inequality and the fact that $\chi^2(n^2)$ has variance $2n^2$. Now note

$$Y(C, D) = \sum_{h=1}^{r} \sum_{i \in C_h} \sum_{j \in D_h} M_{ij} \sim \mathcal{N}(0, rK^2)$$

Standard gaussian tail bounds imply that

$$\begin{aligned}\mathbb{P}\left[Y(C,D) > 2rK^{3/2}w\sqrt{(\log n + \log r)}\right] &\le \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2rK^2}\left(2rK^{3/2}w\sqrt{(\log n + \log r)}\right)^2\right) \\ &\le (nr)^{-2rKw^2}\end{aligned}$$

A crude upper bound on the number of pairs (C, D) is

$$\left(\binom{n}{rK} r^{rK}\right)^2 = o\left((nr)^{2rK}\right)$$

and therefore a union bound implies that $s_I(M) = \max_{C,D} Y(C, D) \le 2rK^{3/2}w\sqrt{(\log n + \log r)}$ with probability $1 - o_n(1)$. This completes the proof of the lemma. ■

Proof (of Corollary 94) Consider the following reduction $\mathcal{A}$ that adds a simple post-processing step to k-PDS-TO-GHPM as in Corollary 88. On input graph G with N vertices:

1. Form the graph M_{R} by applying k-PDS-TO-GHPM to G with parameters N, r, k, E, ℓ, n, s and μ where μ is given by

$$\mu = \frac{r^t\sqrt{r}}{(r-1)} \cdot \Phi^{-1}\left(\frac{1}{2} + \frac{1}{2} \cdot \min\{P_0, 1-P_0\}^{-1} \cdot \gamma\right)$$

   and Φ^{-1} is the inverse of the standard normal CDF.

2. Let G_1 be the graph where each edge (i,j) with is in G_1 if and only if $(M_{\mathrm{R}})_{ij} \geq 0$. Now form G_2 as in Step 2 of Corollary 88, while restricting to edges between the two parts.

This clearly runs in poly(N) time and it suffices to establish its approximate Markov transition properties. Let $\mathcal{A}_1$ denote the first step with input G and output M_{R}, and let $\mathcal{A}_2$ denote the second step with input M_{R} and output G_2. Let C and D be two fixed sequences, each consisting of r disjoint subsets of $[ksr^t]$ of size kr^{t-1}. Let $P_1, P_2 \in (0,1)$ be

$$P_1 = \Phi\left(\frac{\mu(r-1)}{r^t\sqrt{r}}\right) \quad \text{and} \quad P_2 = \Phi\left(-\frac{\mu}{r^t\sqrt{r}}\right)$$

Note that by the definition of μ, we have that $P_1 = \frac{1}{2} + \frac{1}{2} \cdot \min\{P_0, 1-P_0\}^{-1} \cdot \gamma$. Now note that $\mathcal{A}_2$ applied to $M_{\mathrm{R}} \sim \mathrm{GHPM}_D(ksr^t, r, C, D, \gamma)$ yields an instance of $\mathrm{BHPM}_D(ksr^t, r, C, D, \gamma)$ with the following modified edge probabilities:

1. The edge probabilities between vertices C_h and D_h for each $1 \leq h \leq r$ are still $P_0 + \gamma$.

2. The edge probabilities between C_{h_1} and D_{h_2} for each $h_1 \neq h_2$ are now

$$P_0 + 2\min\{P_0, 1-P_0\} \cdot \left(\Phi\left(-\frac{\mu}{r^t\sqrt{r}}\right) - \frac{1}{2}\right) = P_0 + 2\min\{P_0, 1-P_0\} \cdot \left(P_2 - \frac{1}{2}\right)$$

3. All other edge probabilities are still P_0.

We now apply a similar sequence of inequalities as in Corollary 88. For now assume that $P_0 \leq 1/2$. Using the fact that all of the edge indicators of this model and the usual definition of BHPM are independent, the tensorization property in Fact 15 and Lemma 18, we now have that

$$\begin{aligned}
&d_{\mathrm{TV}}\left(\mathcal{A}_2\left(\mathrm{GHPM}_D(ksr^t, r, C, D, \gamma)\right), \mathrm{BHPM}_D(ksr^t, r, C, D, \gamma)\right) \\
&\quad \leq d_{\mathrm{TV}}\left(\mathrm{Bern}\left(P_0 - \frac{\gamma}{r-1}\right)^{\otimes k^2 r^{2t-1}(r-1)}, \mathrm{Bern}\left(P_0 + 2P_0 \cdot \left(P_2 - \frac{1}{2}\right)\right)^{\otimes k^2 r^{2t-1}(r-1)}\right) \\
&\quad \leq \left|\frac{\gamma}{r-1} + 2P_0 \cdot \left(P_2 - \frac{1}{2}\right)\right| \cdot \sqrt{\frac{k^2 r^{2t-1}(r-1)}{2\left(P_0 - \frac{\gamma}{r-1}\right)\left(1 - P_0 + \frac{\gamma}{r-1}\right)}} \\
&\quad \leq \left|\frac{\gamma}{r-1} + 2P_0 \cdot \left(P_2 - \frac{1}{2}\right)\right| \cdot O\left(kr^t\right)
\end{aligned}$$

where the third inequality uses the fact that P_0 is bounded away from 0 and 1 and $\gamma = o(1)$. Now note that

$$\frac{\gamma}{r-1} = \frac{2P_0}{r-1} \cdot \left(\Phi\left(\frac{\mu(r-1)}{r^t\sqrt{r}} \right) - \frac{1}{2} \right)$$

Using the standard Taylor approximation for $\Phi(x) - 1/2$ around zero when $x \in (-1, 1)$, we have

$$\begin{aligned} \left| \frac{\gamma}{r-1} + 2P_0 \cdot \left(P_2 - \frac{1}{2} \right) \right| &= 2P_0 \cdot \left| \frac{1}{r-1} \left(\Phi\left(\frac{\mu(r-1)}{r^t\sqrt{r}} \right) - \frac{1}{2} \right) - \left(\Phi\left(-\frac{\mu}{r^t\sqrt{r}} \right) - \frac{1}{2} \right) \right| \\ &= O\left(\frac{\mu^3\sqrt{r}}{r^{3t}} \right) \end{aligned}$$

Therefore we have that

$$d_{\mathrm{TV}}\left(\mathcal{A}_2\left(\mathrm{GHPM}_D(ksr^t, r, C, D, \gamma) \right), \mathrm{BHPM}_D(ksr^t, r, C, D, \gamma) \right) = O\left(\frac{k\mu^3\sqrt{r}}{r^{2t}} \right)$$

A nearly identical argument considering the complement of the graph G_1 and replacing with P_0 with $1 - P_0$ establishes this bound in the case when $P_0 > 1/2$. Observe that $\mathcal{A}_2\left(\mathcal{N}(0,1)^{\otimes n \times n}\right) \sim \mathcal{G}_B(n, n, P_0)$. Now consider applying Lemma 16 to the steps $\mathcal{A}_1$ and $\mathcal{A}_2$ as in Corollary 88. It can be verified that the given bound on γ yields the condition on μ needed to apply Theorem 92 if $c > 0$ is sufficiently small. Thus ϵ_1 is bounded by Theorem 92 and ϵ_2 is bounded by the argument above after averaging over C and D and applying the conditioning property of Fact 15. This application of Lemma 16 therefore yields the desired two approximate Markov transition properties and completes the proof of the corollary. ■

Proof (of Theorem 12) As discussed in the beginning of this section, it suffices to map to $\mathcal{G}(n, P_0 - \mu_1)$ under H_0 and $\mathrm{TSI}(n, k, k_1, P_0, \mu_1, \mu_2, \mu_3)$ under H_1 where $\mu_3 = P_1 - P_0$ and $\mu_1, \mu_2 \geq 0$. Thus it suffices to show that the reduction $\mathcal{A}$ in Corollary 98 fills out all of the possible growth rates specified by the computational lower bound $\frac{(P_1-P_0)^2}{P_0(1-P_0)} = \tilde{o}(n/k^2)$ and the other conditions in the theorem statement. Fix a constant pair of probabilities $0 < q < p \leq 1$ and any sequence of parameters (n, k, P_1, P_0) all of which are implicitly functions of n such that

$$\frac{(P_1 - P_0)^2}{P_0(1 - P_0)} \leq c \cdot \frac{n}{w^3 \cdot k^2 \log n} \quad \text{and} \quad \min\{P_0, 1 - P_0\} = \Omega_n(1)$$

for sufficiently large n, sufficiently small constant $c > 0$ and an arbitrarily slow-growing increasing positive integer-valued function $w = w(n) \to \infty$ at least satisfying that $w(n) = n^{o(1)}$. As in the proof of Theorem 4, it suffices to specify:

1. a sequence (N, k_N) such that the k-$\mathrm{PDS}(N, k_N, p, q)$ is hard according to Conjecture 3; and
2. a sequence $(n', k', P_1, P_0, s, t, \mu)$ satisfying: (2.1) the parameters (n', k', P_1, P_0) are in the regime of the desired computational lower bound for SEMI-CR; (2.2) (n', k') have the same growth rates as (n, k); and (2.3) such that $\mathcal{G}(n', P_0 - \mu_1)$ and $\mathrm{TSI}(n', k', k'/2, P_0, \mu_1, \mu_2, P_1 - P_0)$, where k' is even and $\mu_1, \mu_2 \geq 0$, can be produced by $\mathcal{A}$ with input k-$\mathrm{PDS}(N, k_N, p, q)$.

We choose these parameters as follows:

- let t be such that 3^t is the smallest power of 3 larger than $k/\sqrt{n}$ and let $s = \lceil 2n/3k \rceil$;

- let $\mu \in (0,1)$ be given by

$$\mu = 3^t \cdot \Phi^{-1}\left(\frac{1}{2} + \frac{1}{2} \cdot \min\{P_0, 1-P_0\}^{-1}(P_1 - P_0)\right)$$

- now let

$$k_N = \left\lfloor \frac{1}{2}\left(1 + \frac{p}{Q}\right)^{-1} w^{-2} \cdot \sqrt{n} \right\rfloor$$

  where $Q = 1 - \sqrt{(1-p)(1-q)} + \mathbf{1}_{\{p=1\}}\left(\sqrt{q} - 1\right)$; and

- let $n' = 3k_N s \cdot \frac{3^t - 1}{2}$, let $k' = (3^t - 1)k_N$ and let $N = wk_N^2$.

Note that $3^t = \Theta(k/\sqrt{n})$, $s = \Theta(n/k)$ and $3^t k_N s \leq \text{poly}(N)$. Note that this choice of μ implies that

$$P_1 = P_0 + 2\min\{P_0, 1-P_0\} \cdot \left(\Phi\left(\frac{\mu}{3^t}\right) - \frac{1}{2}\right)$$

which implies that the instance of TSI output by $\mathcal{A}$ has edge density P_1 on its k'-vertex the planted dense subgraph. It follows that

$$n' \asymp 3^t k_N s \asymp \frac{k}{\sqrt{n}} \cdot \frac{n}{k} w^{-2} \cdot \sqrt{n} \asymp w^{-2} \cdot n \quad \text{and} \quad k' \asymp 3^t k_N \asymp w^{-2} k$$

$$\frac{(P_1 - P_0)^2}{P_0(1-P_0)} \leq c \cdot \frac{n}{w^3 \cdot k^2 \log n} \lesssim c \cdot \frac{n'}{w \cdot (k')^2 \log n'}$$

$$m \leq 2\left(\frac{p}{Q} + 1\right) wk_N^2 \leq w^{-1}\sqrt{n} \cdot k_N \leq 3^t k_N s$$

$$\mu \lesssim 3^t \cdot (P_1 - P_0) \lesssim 3^t \cdot \frac{\sqrt{n}}{w^{3/2} \cdot k\sqrt{\log n'}} \leq \frac{c}{w^{3/2}\sqrt{\log n'}}$$

where the last bound above follows from the fact that $\Phi(x) - 1/2 \sim x$ if $|x| \to 0$. Here, m is the smallest multiple of k_N larger $\left(\frac{p}{Q} + 1\right) N$. Now note that: (2.1) the third inequality above on $(P_1 - P_0)^2/P_0(1-P_0)$ implies that (n', k', P_1, P_0) is in the desired hard regime; (2.2) (n, n') and (k, k') have the same growth rates since $w = n^{o(1)}$; and (2.3) the last two bounds above imply that taking c small enough yields the conditions needed to apply Corollary 98 to yield the desired reduction. This completes the proof of the theorem. ■

Proof (of Lemma 103) The parameters a, μ_1, μ_2 for which these distributional statements are true are given by

$$a = \Phi(\tau) - \Phi(-\tau)$$

$$\mu_1 = \frac{1}{2}\left((1 - \Phi(\tau - \mu)) - \Phi(-\tau - \mu)\right) = \frac{1}{2}\left(\Phi(\tau + \mu) - \Phi(\tau - \mu)\right)$$

$$\mu_2 = \frac{1}{2}\left(\Phi(\tau) - \Phi(-\tau)\right) - \frac{1}{2}\left(\Phi(\tau + \mu) - \Phi(-\tau + \mu)\right) = \frac{1}{2}\left(2 \cdot \Phi(\tau) - \Phi(\tau + \mu) - \Phi(\tau - \mu)\right)$$

Now note that

$$\mu_1 = \frac{1}{2}\left(\Phi(\tau + \mu) - \Phi(\tau - \mu)\right) = \frac{1}{2\sqrt{2\pi}} \int_{\tau - \mu}^{\tau + \mu} e^{-t^2/2} dt = \Theta(\mu)$$

and is positive since $e^{-t^2/2}$ is bounded on $[\tau-\mu, \tau+\mu]$ as τ is constant and $\mu \to 0$. Furthermore, note that

$$\mu_2 = \frac{1}{2}\left(2 \cdot \Phi(\tau) - \Phi(\tau+\mu) - \Phi(\tau-\mu)\right) = \frac{1}{2\sqrt{2\pi}} \int_{\tau-\mu}^{\tau} e^{-t^2/2} dt - \frac{1}{2\sqrt{2\pi}} \int_{\tau}^{\tau+\mu} e^{-t^2/2} dt$$
$$= \frac{1}{2\sqrt{2\pi}} \int_{\tau}^{\tau+\mu} \left(e^{-(t-\mu)^2/2} - e^{-t^2/2}\right) dt = \frac{1}{2\sqrt{2\pi}} \int_{\tau}^{\tau+\mu} e^{-t^2/2} \left(e^{t\mu-\mu^2/2} - 1\right) dt$$

Now note that as $\mu \to 0$ and for $t \in [\tau, \tau+\mu]$, it follows that $0 < e^{t\mu-\mu^2/2} - 1 = \Theta(\mu)$. This implies that $0 < \mu_2 = \Theta(\mu^2)$, as claimed. ■

Proof (of Theorem 11) To prove this theorem, we will to show Theorem 104 implies that k-BPDS-TO-GLSM fills out all of the possible growth rates specified by the computational lower bound $n = \tilde{o}\left(\tau_{\mathcal{U}}^{-4}\right)$ and the other conditions in the theorem statement, as in the proof of Theorems 4 and 82. Fix a constant pair of probabilities $0 < q < p \leq 1$ and any sequence $(n, k, d, \mathcal{U})$ where $\mathcal{U} = (\mathcal{D}, \mathcal{Q}, \{\mathcal{P}_\nu\}_{\nu \in \mathbb{R}}) \in \text{UC}(n)$ all of which are implicitly functions of n with

$$n \leq \frac{c}{\tau_{\mathcal{U}}^4 \cdot w^2 \cdot (\log n)^2} \quad \text{and} \quad wk^2 \leq d$$

for sufficiently large n, an arbitrarily slow-growing function $w = w(n) \to \infty$ and a sufficiently small constant $c > 0$. Now consider specifying the parameters M, N, k_M, k_N and t exactly as in Theorem 82. Now note that under these parameter settings, we have that

$$\tau_{\mathcal{U}} \leq \frac{c^{1/4}}{n^{1/4} w^{1/2} \sqrt{\log n}} \leq 2c^{1/4} \cdot \sqrt{\frac{k_N}{N \log N}}$$

Therefore $\tau_{\mathcal{U}}$ satisfies the conditions needed to apply Theorem 104 for a sufficiently small $c > 0$. The other parameters $(n, k, d, \mathcal{U})$ and (M, N, k_M, k_N, p, q) can also be verified to satisfy the conditions of this theorem. We now have that k-BPDS(M, N, k_M, k_N, p, q) is hard according to Conjecture 3, and that GLSM$(n, k, d, \mathcal{U})$ can be produced by the reduction k-BPDS-TO-GLSM applied to BPDS(M, N, k_M, k_N, p, q). This verifies the criteria in Condition E.1 and, following the argument in Section E.2, Lemma 14 now implies the theorem. ■