One fundamental goal of high-dimensional statistics is to detect or recover planted structure (such as a low-rank matrix) hidden in noisy data. A growing body of work studies low-degree polynomials as a restricted model of computation for such problems: it has been demonstrated in various settings that low-degree polynomials of the data can match the statistical performance of the best known polynomial-time algorithms. Prior work has studied the power of low-degree polynomials for the task of detecting the presence of hidden structures. In this work, we extend these methods to address problems of estimation and recovery (instead of detection). For a large class of “signal plus noise” problems, we give a user-friendly lower bound for the best possible mean squared error achievable by any degree-$D$ polynomial. To our knowledge, these are the first results to establish low-degree hardness of recovery problems for which the associated detection problem is easy. As applications, we give a tight characterization of the low-degree minimum mean squared error for the planted submatrix and planted dense subgraph problems, resolving (in the low-degree framework) open problems about the computational complexity of recovery in both cases.

1. Introduction. Many problems in high-dimensional statistics exhibit a gap between what is achievable statistically and what is achievable with known computationally-efficient (i.e., polynomial-time) algorithms. Such information-computation gaps appear in many of the canonical models of statistical estimation problems, including sparse principal component analysis (PCA), planted clique, and community detection, among others. Because these are average-case problems in which the input is drawn from a specially chosen probability distribution, it is unlikely that the computational hardness of such problems can be established under standard worst-case complexity assumptions such as $P \neq NP$ (e.g. [48, 24, 2]). Instead, to provide rigorous evidence for such gaps, researchers either give reductions between different statistical problems (e.g. [21, 78, 29]), or prove lower bounds in restricted models of computation; these include lower bounds against families of convex programs (see e.g. [88]), lower bounds in the statistical query framework (e.g. [71, 49]), lower bounds against local algorithms (e.g. [55, 18]), and more.

The focus of this work is the low-degree polynomial model of computation, in which we require that our algorithm’s output is computable via a polynomial of bounded degree in the input. This model has recently come into focus as a promising framework for studying the complexity of hypothesis testing problems. The study of low-degree polynomials for hypothesis testing was initiated implicitly in the work of Barak et al. on sum-of-squares lower bounds for planted clique [15], and was subsequently refined and extended to numerous additional settings by Hopkins and Steurer [68], followed by others (see e.g. [64, 63], and also [74] for a survey). In these works, the goal is to hypothesis test (with asymptotically vanishing error probability) between a null distribution (typically i.i.d. “noise”) and
a planted distribution (which includes a planted structure hidden in noise). Many state-of-the-art algorithms for such problems—including spectral methods and approximate message passing (AMP) algorithms [46]—can be represented as low-degree (multivariate) polynomial functions of the input, where “low” means logarithmic in the dimension. Furthermore, it has been shown that the class of low-degree polynomials is precisely as powerful as the best known polynomial-time algorithms for many canonical problems, including planted clique [15], sparse PCA [44], community detection [68], tensor PCA [64, 74], and others. From this picture emerges the intriguing conjecture that low-degree polynomials may be as powerful as any polynomial-time algorithm for a broad class of high-dimensional testing problems [63]. Thus, an impossibility result for low-degree polynomials is not merely a lower bound within a restricted model of computation, but further constitutes compelling evidence for average-case computational hardness.

While the low-degree framework has had many successes, one limitation of the existing theory is that it is restricted to hypothesis testing (also called detection) problems (although one exception is the work of [53, 97, 30], which studies low-degree polynomials in the context of random optimization problems with no planted signal). But more often, in high-dimensional statistics we are interested in recovery or estimation, where the goal is to approximate the planted structure (in some norm, which may vary with the application) rather than merely detect its presence. For some problems (e.g. planted clique) detection and recovery are believed to be equally hard in the sense that both tasks admit polynomial-time algorithms in precisely the same regime of parameters. In such cases, computational hardness of recovery can often be deduced from computational hardness of detection (via a polynomial-time reduction from detection to recovery, as in Section 5.1 of [78]). On the other hand, other problems are believed to exhibit detection-recovery gaps where the recovery task is strictly harder (computationally) than the associated detection task. For such problems, existing work has often struggled to find compelling concrete evidence for hardness of recovery in the parameter regime where detection is easy. In this work, we refer to detection-recovery gaps as situations where the computational limits of detection and recovery differ; there are also situations where the statistical limits of detection and recovery differ (e.g. [45, 80]).

One popular problem that appears to exhibit a detection-recovery gap is the following planted submatrix problem (studied by e.g. [72, 31, 32, 78, 38, 33, 60, 52] and also used as a model for sparse PCA in the spiked Wigner model [42, 16, 19]), where a submatrix of elevated mean is hidden in a Gaussian random matrix. In the planted submatrix problem, we observe an $n \times n$ matrix $Y = \lambda vv^\top + W$ where $\lambda > 0$ is the signal-to-noise ratio (SNR), $v \in \{0, 1\}^n$ is a planted signal with i.i.d. Bernoulli($\rho$) entries, and $W$ is a symmetric matrix of Gaussian $\mathcal{N}(0, 1)$ noise (see Definition 2.4). We are interested in the high-dimensional setting, where $n \to \infty$ with $\lambda = n^{-a}$ and $\rho = n^{-b}$ for constants $a > 0$ and $0 < b < 1$. (To see why this is the interesting regime for $a$ and $b$: if $a < 0$ then recovery is easy by entrywise thresholding, and if $b > 1$ then the planted submatrix typically has size zero.) When $b < 1/2$ there appears to be a detection-recovery gap: distinguishing between $Y = \lambda vv^\top + W$ and the null distribution $Y = W$ is easy when $a < 2(1/2 - b)$, simply by summing all entries of $Y$; however, when $a > 1/2 - b$ there are no known polynomial-time algorithms for recovering $v$, or even for producing a non-trivial estimate of $v$. (The reader may wonder whether the detection-recovery gap can be closed by simply choosing a better null distribution that matches the mean and covariance of the planted distribution. We show in Appendix B in the Supplement that this closes the gap partially but not all the way: detection is still easy when $a < \frac{1}{3}(1/2 - b)$.)

In this work we give the first results that directly address recovery (as opposed to detection) in the low-degree framework. Suppose we are given some observation $Y \in \mathbb{R}^N$, and the goal
is to estimate a scalar quantity $x \in \mathbb{R}$ (which could be, for instance, the first coordinate of the signal vector). Let $\mathbb{R}[Y]_{\leq D}$ denote the space of polynomials $f : \mathbb{R}^N \rightarrow \mathbb{R}$ of degree at most $D$. Define the degree-$D$ minimum mean squared error

$$\text{MMSE}_{\leq D} := \inf_{f \in \mathbb{R}[Y]_{\leq D}} \mathbb{E}(f(Y) - x)^2$$

where the expectation is over the joint distribution $\mathbb{P}$ of $x$ and $Y$. No generality is lost by restricting to polynomials with deterministic (as opposed to random) coefficients; see Appendix A in the Supplement. As we see below, understanding MMSE$_{\leq D}$ is equivalent to understanding the degree-$D$ maximum correlation (which will be technically more convenient)

$$\text{Corr}_{\leq D} := \sup_{f \in \mathbb{R}[Y]_{\leq D}} \mathbb{E}(f(Y) \cdot x) = \sup_{f \in \mathbb{R}[Y]_{\leq D}} \frac{\mathbb{E}_P[f(Y) \cdot x]}{\sqrt{\mathbb{E}_P[f(Y)^2]}}.$$

The expert reader will note that in contrast to prior low-degree polynomial lower bounds (e.g. [68]), there is no null distribution involved in this expression; the expectations in the numerator and denominator are both over the planted distribution $\mathbb{P}$. From now on, expectations will be implicitly taken over $\mathbb{P}$ unless stated otherwise.

**Fact 1.1.** $\text{MMSE}_{\leq D} = \mathbb{E}[x^2] - \text{Corr}_{\leq D}^2$.

**Proof.** Suppressing the constraint $f \in \mathbb{R}[Y]_{\leq D}$ for ease of notation, we have

$$\text{MMSE}_{\leq D} = \inf_{\mathbb{E}[f^2] = 1} \inf_{\alpha \in \mathbb{R}} \mathbb{E}(\alpha f(Y) - x)^2 = \inf_{\mathbb{E}[f^2] = 1} \mathbb{E}[x^2] - \mathbb{E}[f(Y) \cdot x]^2$$

$$= \mathbb{E}[x^2] - \left( \sup_{\mathbb{E}[f^2] = 1} \mathbb{E}[f(Y) \cdot x] \right)^2 = \mathbb{E}[x^2] - \text{Corr}_{\leq D}^2,$$

completing the proof.

For the case of the planted submatrix problem (discussed above) we will take $x$ to be the first coordinate of the signal: $x = v_1$. Note that due to symmetry, recovering $v_1$ is equivalent to recovering the entire vector $v$ in the sense that

$$\inf_{f_1, \ldots, f_n \in \mathbb{R}[Y]_{\leq D}} \frac{1}{n} \sum_{i=1}^n (f_i(Y) - v_i)^2 = \inf_{f \in \mathbb{R}[Y]_{\leq D}} \mathbb{E}(f(Y) - v_1)^2 = \text{MMSE}_{\leq D}.$$

We remark that in some problems, inherent symmetries make it impossible to discern whether the planted vector is $v$ or $-v$. In such cases, one can break symmetry by choosing $x = v_1 v_2$. However, this issue does not arise for the examples considered in this paper.

### 1.1. Our Contributions.

While MMSE$_{\leq D}$ is a natural quantity, in many cases it is difficult to bound, and for this reason it has not yet received attention in prior work. In this work, we obtain the first useful lower bounds on MMSE$_{\leq D}$ in various “signal plus noise” settings, namely the general additive Gaussian noise model (see Section 2.1) and the general binary observation model (see Section 2.3). This allows us to tightly characterize MMSE$_{\leq D}$ for both the planted submatrix problem (see Section 2.2) and the related planted dense subgraph problem (see Section 2.4). For example, we show the following for planted submatrix (restricting to the most interesting regime $b < 1/2$, where there appears to be a detection-recovery gap).
THEOREM 1.2 (Special case of Theorem 2.5). Consider the planted submatrix problem with $n \to \infty$, $\lambda = n^{-a}$, and $\rho = n^{-b}$ for constants $a > 0$ and $0 < b < 1/2$.

(i) If $a > 1/2 - b$ then $\text{MMSE}_{\leq n^c} = \rho - (1 + o(1))\rho^2$ for some constant $\epsilon = \epsilon(a, b) > 0$, i.e., no degree-$n^c$ polynomial outperforms the trivial estimator $f(Y) = \mathbb{E}[v_1] = \rho$ (which has mean squared error $\rho - \rho^2$).

(ii) If $a < 1/2 - b$ then $\text{MMSE}_{\leq C} = o(\rho)$ for some constant $C = C(a, b)$, i.e., some degree-$C$ polynomial achieves asymptotically perfect estimation.

Part (ii) is unsurprising and simply confirms that low-degree polynomials succeed in the regime where recovery is already known to be computationally easy. Our main result is part (i), which establishes that low-degree polynomials cannot be used for recovery throughout the “hard” regime where no computationally-efficient algorithms are known. This resolves (in the low-degree framework) an open problem that has been mentioned in various works [9, 78, 38]. Crucially, our result shows hardness of recovery in a regime where detection is easy, and thus provides concrete evidence for a detection-recovery gap. There is nothing fundamental about the choice of degree $n^c$ in part (i); this is simply the highest degree we are able to rule out. As discussed in Section 1.1.1 below, even ruling out degree $O(\log n)$ is considered evidence that no polynomial-time algorithm exists.

While Theorem 2.5 focuses on the goal of estimating $v$ in $\ell_2$ loss, another natural task is support recovery (also called localization), where the goal is to estimate $v$ in Hamming ($\ell_0$) loss. In Appendix J we show that these notions are equivalent for our purposes: if there is a polynomial-time estimator $\hat{v}$ achieving $\|\hat{v} - v\|_2^2 = o(n\rho)$ then there is a polynomial-time estimator $\hat{u}$ achieving $\|\hat{u} - v\|_0 = o(n\rho)$, and vice versa. Thus, part (i) above suggests that support recovery is also hard when $a > 1/2 - b$ (although we have not shown unconditionally that low-degree polynomials fail at support recovery). On the positive side, the proof of part (ii) actually shows that when $a < 1/2 - b$, it is possible to exactly recover $v$ with high probability by thresholding a low-degree polynomial (see Appendix I).

1.1.1. Implications for other models of computation. We note that many of the best known algorithmic approaches can be represented as (or approximated by) polynomials of degree $O(\log n)$, and so part (i) of Theorem 2.5 implies failure of any such algorithm in the “hard” regime. One such family of algorithms are spectral methods which involve computing the leading eigenvector of some symmetric matrix $M$ (of dimension $\text{poly}(n)$) constructed from the data $Y$. The matrix $M$ can either be $Y$ itself (e.g. [7, 50, 20]), or a more sophisticated function of $Y$ (e.g. [67, 66, 64, 65]). It is typical for the leading “signal” eigenvalue to be larger than the rest by a constant factor, in which case $O(\log n)$ rounds of power iteration suffice to approximate the leading eigenvector. If each entry of $M$ is a constant-degree polynomial in $Y$ (which is the case for most natural spectral algorithms), the whole process amounts to computing an $O(\log n)$-degree polynomial. Another family of low-degree algorithms are those based on the approximate message passing (AMP) framework [46] (see also e.g. [25, 17, 90, 69]). These typically involve a constant number of nonlinear iterations, each of which can be well-approximated by a constant-degree polynomial; thus the whole process is a constant-degree polynomial. In some cases, a spectral initialization is needed (e.g. [83]), bringing the total degree to $O(\log n)$. Finally, other state-of-the-art algorithms have been designed directly from low-degree polynomials [68, 14]. Part (i) of Theorem 2.5 rules out the success of all of these algorithms; in fact, it also rules out polynomials of much higher degree: $n^{\Omega(1)}$ instead of merely $O(\log n)$, which suggests that the runtime required is $\exp(n^{\Omega(1)})$ instead of merely super-polynomial (see Hypothesis 2.1.5 of [63], or [74, 44]).
1.2. Prior Work. In this section, we discuss how low-degree polynomials compare to some other restricted models of computation which are popular in the study of information-computation gaps. While each model offers valuable (and complementary) insights, we emphasize that our work is the only one to simultaneously meet the following two criteria:

- Our approach directly addresses the recovery problem and is able to establish detection-recovery gaps, i.e., it can show hardness of recovery in regimes where detection is easy.
- Our approach rules out a class of algorithms (namely low-degree polynomials) which are as powerful as all known polynomial-time algorithms for standard testbed problems such as planted clique, sparse PCA, and tensor PCA. (As we will see, some methods do not predict the correct computational thresholds for these problems, at least without some caveats.)

For the specific applications we consider—planted submatrix and planted dense subgraph—we defer an in-depth discussion of prior work to Sections 2.2 and 2.4, respectively.

1.2.1. Low-degree likelihood ratio. In the setting of hypothesis testing between a planted distribution $P_n$ and a null distribution $Q_n$ (where $n$ is a notion of problem size, e.g. dimension), the most closely related prior work (including [68, 64, 63, 11, 74, 44]) has analyzed (the norm of) the low-degree likelihood ratio

$$
\|L^{\leq D}\| := \sup_{f \in \mathbb{R}^{|Y|} \subseteq D} \frac{\mathbb{E}_{Y \sim P_n}[f(Y)]}{\sqrt{\mathbb{E}_{Y \sim Q_n}[f(Y)]^2}}.
$$

The notation $\|L^{\leq D}\|$ stems from the fact that this quantity can be computed by projecting the likelihood ratio $L_n = \frac{dP_n}{dQ_n}$ onto the subspace of degree-$D$ polynomials (see e.g. [63, 74] for details). As is made clear in e.g. [63], we may equivalently define

$$
\|L^{\leq D}\|^2 - 1 = \left( \sup_{f \in \mathbb{R}^{|Y|} \subseteq D} \frac{\mathbb{E}_{Y \sim P_n}[f(Y)]}{\sqrt{\text{Var}_{Y \sim Q_n}[f(Y)]}} \right)^2.
$$

The quantity $\|L^{\leq D}\|$ is a heuristic measure of how well degree-$D$ polynomials can distinguish $P_n$ from $Q_n$. To see why, note that the variational problem in (1.4) relaxes a program which optimizes over degree-$D$ polynomials $f$ that test between $Q_n$ and $P_n$: if $\|L^{\leq D}\| = O(1)$ then no degree-$D$ polynomial $f$ can separate $P_n$ from $Q_n$ in the sense $\mathbb{E}_{P_n}[f] - \mathbb{E}_{Q_n}[f] \geq 1$, $\text{Var}_{Q_n}[f] = o(1)$, and $\text{Var}_{P_n}[f] = o(1)$.

The minimum degree of a polynomial hypothesis test (and $\|L^{\leq D}\|$ as a proxy for such) is an interesting measure of problem complexity in its own right—the idea of studying a function’s degree as a proxy for its computational complexity has a rich history in the theory of worst-case complexity, e.g. [84, 86]. Furthermore, a polynomial hypothesis test of degree $D$ (whose coefficients have polynomial bit complexity) can always be implemented by an algorithm with running time $O(n^D)$, as one may simply take the weighted sum of the values of $n^D$ monomials. Some polynomials with special structure, such as those corresponding to spectral algorithms in matrices with a spectral gap, can be implemented even more efficiently—there, a degree-$O(\log n)$ polynomial can be implemented in time $n^{O(1)}$ using power iteration. A priori, it is unclear that the failure low-degree polynomial hypothesis tests should imply anything for other, more powerful models of computation. But remarkably, the behavior of $\|L^{\leq D}\|$ for $D \approx \log n$ has been observed to coincide with the conjectured computational threshold in many detection problems, including planted clique, community detection, tensor PCA, and sparse PCA [15, 68, 64, 63, 74, 44]: $\|L^{\leq D}\| \to \infty$ as $n \to \infty$ in the “easy” regime (where polynomial-time algorithms are known) while $\|L^{\leq D}\| = O(1)$ in the “hard” regime.
It has been conjectured [63] that the boundedness of \( \|L^D\| \) as the dimension goes to infinity indicates computational intractability for a large class of high-dimensional testing problems. (See [63, 74, 61] for discussion regarding the class of problems for which the low-degree conjecture is believed to hold.) Assuming this conjecture, bounding \( \|L^D\| \) for any \( D \) with \( \frac{D}{\log n} \to \infty \) as \( n \to \infty \) implies hardness of detection. However, this approach is limited to regimes where detection is hard, and so cannot be used to establish detection-recovery gaps.

### 1.2.2. Null-normalized correlation

Hopkins and Steurer [68], in one of the early works that proposed the low-degree likelihood ratio, also proposed a related heuristic for understanding recovery problems. Namely, they study the quantity

\[
\sup_{f \in \mathbb{R}^n : \|f\| \leq D} \frac{\mathbb{E}_{Y \sim P_n}[f(Y) \cdot x]}{\sqrt{\mathbb{E}_{Y \sim Q_n}[f(Y)^2]}}.
\]

Note that this resembles the maximum correlation \( \text{Corr}^D \) defined in (1.2) (which is related to \( \text{MMSE}^D \) via Fact 1.1), except the expectation in the denominator is not taken with respect to the planted distribution but rather with respect to some appropriate choice of null distribution—this is purely a matter of convenience, as it makes bounding (1.5) analytically tractable when \( Q_n \) has independent coordinates. In [68] it is shown that the behavior of (1.5) correctly captures the conjectured computational threshold in the stochastic block model (which does not have a detection-recovery gap). However, we show in Appendix C (in the Supplement) that in the presence of a detection-recovery gap, (1.5) actually captures the detection threshold instead of the recovery threshold: when detection is easy, a polynomial can “cheat” by outputting a much larger value under \( P_n \) than under \( Q_n \), causing (1.5) to diverge to infinity. In this work, we give the first techniques for bounding the more natural quantity \( \text{Corr}^D \), which resolves an open problem of Hopkins and Steurer [68].

### 1.2.3. Sum-of-squares lower bounds

The sum-of-squares (SoS) hierarchy is a powerful family of semidefinite programs that has been widely successful at obtaining state-of-the-art algorithmic guarantees for many problems; see e.g. [88] for a survey. SoS is most naturally suited to the task of certification (or refutation): when there is no hidden structure, SoS can certify (or fail to certify) the absence of structure. SoS lower bounds show that SoS fails to certify the absence of structure (e.g. [15, 73, 64]), providing strong evidence of computational intractability for certification. (In fact, the low-degree likelihood ratio was originally motivated by its connection to the pseudocalibration approach to SoS lower bounds [15, 64, 73].)

However, recent work reveals that the certification problem can sometimes be fundamentally harder than the associated recovery problem (see e.g. [11, 10]); for instance, this seems to be the case for the Sherrington–Kirkpatrick spin glass model [11, 81, 58, 82] and the planted coloring model [12, 10]. Thus, SoS certification lower bounds are not necessarily evidence for computational hardness of recovery. A variant of the SoS hierarchy called the local statistics hierarchy [13] was recently proposed to directly address detection, but there is not currently an analogue for recovery.

### 1.2.4. Spectral algorithms

Spectral algorithms involve computing eigenvectors or singular vectors of matrices constructed from the data; for example, the leading algorithms for a number of PCA tasks are spectral methods [7, 50, 20]. Recently, a line of work has demonstrated that sum-of-squares algorithms for recovery can often be translated into “low-degree” spectral algorithms, in which the resulting matrix has entries that are constant-degree polynomials of the data and constant spectral gap [67, 66, 65]. As discussed in Section 1.1.1 above, such spectral methods can be approximated by low-degree polynomials. Thus, remarkably,
for many recovery problems, even algorithms based on powerful convex programs are equivalent in power to low-degree algorithms. (There is a result due to Hopkins et al. [64] which formally equates the power of SoS and low-degree spectral algorithms for detection in a wide variety of noise-robust problems; however this result does not guarantee that the low-degree matrices have a sufficiently large spectral gap, and so it falls short of implying that SoS is captured by low-degree polynomials.)

1.2.5. **Statistical query algorithms.** For settings where the observed data consists of i.i.d. samples drawn from some distribution, the statistical query (SQ) model [71, 49] is used to understand information-computation tradeoffs. A statistical query algorithm is allowed to compute the average of any (not necessarily efficiently computable) bounded function or “query” on the samples, up to an adversarial error of bounded magnitude—the number of queries is a proxy for computational efficiency, and the magnitude of the error is a proxy for the signal-to-noise ratio. In some cases, it is possible to map a single-sample problem (like the ones we consider in this paper) to a computationally equivalent multi-sample problem, as was done in e.g. [47, 28]. In some cases, predictions in the SQ framework match the conjectured computational thresholds; in other cases, there are discrepancies. For instance, in the case of tensor PCA, SQ lower bounds suggest that there is a detection-recovery gap whereas in reality there is not (i.e., efficient algorithms for recovery are known in the regime where no SQ algorithms using the “VSTAT” oracle exist) [47]. In a concurrent work, [28] compare SQ and low-degree algorithms, showing that lower bounds against both algorithms agree for a large class of detection problems; we refer the reader there for a more thorough discussion.

1.2.6. **Approximate message passing.** The approximate message passing (AMP) framework (e.g. [25, 46, 17, 90, 69]) gives state-of-the-art algorithmic guarantees for a wide variety of problems. In some settings, AMP is information-theoretically optimal (e.g. [41]), and when it is not, AMP is often conjectured to be optimal among polynomial-time algorithms (or at least among nearly-linear-time algorithms) [75, 76]. For this reason, the failure of AMP is often taken as evidence that no efficient algorithm exists (specifically for the recovery problem). However, there are some natural problems—including planted clique and tensor PCA—where AMP is known to have strictly worse performance than other polynomial-time algorithms [43, 91]. There are state-of-the-art algorithms for tensor PCA which can be interpreted as a “lifting” of AMP in some sense [98], but it remains unclear whether similar liftings can be performed more generally (e.g. for planted clique).

AMP (and its liftings) can often be approximated by low-degree polynomials: AMP analyses typically consider a constant number of iterations, and the nonlinearities applied in each step are often well-approximated by constant-degree polynomials (in fact, the use of constant-degree polynomial approximation is important in some analyses; see e.g. [43, 60]). Hence, the success of AMP in such settings is ruled out by our lower bounds. We note that the more traditional analysis of AMP based on the so-called state evolution (SE) equations [46, 17, 69] typically gives sharper results than what we can achieve with our methods; namely, the SE equations allow for the exact mean squared error of AMP (in the limit $n \to \infty$) to be calculated in many cases.

1.2.7. **Optimization landscape.** There are a number of related approaches for understanding the “difficulty” of combinatorial or non-convex optimization landscapes. Typically, one characterizes structural properties of the solutions space, including variants of the overlap gap property (OGP), to prove that certain classes of algorithms fail (e.g. [1, 54, 89, 40, 36, 51, 53]). Restricting our discussion to the context of planted problems, the algorithms ruled out include certain local algorithms and MCMC (Markov chain Monte Carlo)
methods [55, 56, 52, 19]. This framework complements the low-degree framework because low-degree methods do not rule out local/MCMC algorithms and OGP-based methods (for planted problems) do not rule out low-degree algorithms. Like AMP, the OGP approach directly addresses the recovery problem but suffers from one caveat: in some settings—including planted clique and tensor PCA—the natural local and MCMC methods perform strictly worse than the best known polynomial-time algorithms (see [18, 56]). This can sometimes be overcome with problem-specific fixes (e.g. [56, 23, 19]), but there is currently no general framework for lower bounds which captures all plausible fixes.

1.2.8. **Average-case reductions.** Average-case reductions (e.g. [21, 78, 37, 59, 96, 57, 33, 95, 99, 34, 29, 26, 27]) provide fully rigorous connections between average-case problems, showing that if one problem can be solved in polynomial time then so can another. Existing reductions that establish hardness of average-case problems need to start from an average-case problem that is assumed to be hard (such as planted clique or its “secret leakage” variants [27]). Methods such as the low-degree framework complement these results by giving concrete evidence that the starting problem is hard. Existing average-case reductions have only established the presence of detection-recovery gaps when using a starting problem that is assumed to already have a detection-recovery gap. The work of [33] does establish the hardness of recovery for one version of the planted submatrix problem assuming hardness of detection for planted clique; however, the noise in their planted submatrix problem is more complex than the canonical i.i.d. Gaussian noise that we consider here. The work [29] establishes detection-recovery gaps in a few different problems assuming the *planted dense subgraph (PDS) recovery conjecture*, which asserts the presence of a detection-recovery gap in the planted dense subgraph problem. However, concrete evidence for this conjecture has been somewhat lacking until now: in this paper, we give tight recovery lower bounds for planted dense subgraph in the low-degree framework.

### Proof Techniques

We now summarize the difficulties in proving lower bounds on $\text{MMSE}_{\leq D}$, and explain the key insights that we use to overcome these. By Fact 1.1, it is equivalent to instead prove an upper bound on $\text{Corr}_{\leq D}$, defined in (1.2). Suppose we choose a basis $\{h_\alpha\}$ for $\mathbb{R}[Y]_{\leq D}$, which could be, for instance, the standard monomial basis or the Hermite basis. Expanding an arbitrary polynomial $f \in \mathbb{R}[Y]_{\leq D}$ as $f = \sum_\alpha \hat{f}_\alpha h_\alpha$ with coefficients $\hat{f}_\alpha \in \mathbb{R}$, and treating $\hat{f} = (\hat{f}_\alpha)$ as a vector, we can equivalently express $\text{Corr}_{\leq D}$ as

$$\text{Corr}_{\leq D} = \sup_{\hat{f}} \frac{\langle c, \hat{f} \rangle}{\sqrt{\hat{f}^\top P \hat{f}}}$$

where

$$c_\alpha := \mathbb{E}[h_\alpha(Y) \cdot x]$$

and

$$P_{\alpha\beta} := \mathbb{E}[h_\alpha(Y) h_\beta(Y)].$$

We can compute $c$ and $P$ explicitly, and then after the change of variables $g = P^{1/2} \hat{f}$ we have that the value of (1.6) is

$$\text{Corr}_{\leq D} = \sup_g \frac{c^\top P^{-1/2} g}{\|g\|} = \sqrt{c^\top P^{-1} c},$$
since the optimizer is $g = P^{-1/2} c$. Thus, we have an explicit formula for $\text{Corr}_{\leq D}$. However, it seems difficult to analytically control this expression due to the matrix inversion $P^{-1}$ (which in most examples does not seem tractable to express in closed form). If $\{h_\alpha\}$ were a basis of orthogonal polynomials (with respect to the planted distribution) then $P$ would be a diagonal matrix and inversion would be trivial; however, it is not clear how to find such an orthogonal basis in closed form.

We are able to overcome the above difficulties by performing some simplifying manipulations: we apply Jensen’s inequality to the “signal” but not the “noise”. Concretely, consider the additive Gaussian noise model $Y = X + Z$ where the signal $X$ is drawn from some prior, and $Z$ is i.i.d. Gaussian noise. In this case, we can bound the denominator of (1.2) using

$$\mathbb{E}[f(Y)^2] = \mathbb{E}_Z \mathbb{E}_X \mathbb{E}_Y f(X + Z)^2 \geq \mathbb{E}_Z \left( \mathbb{E}_X f(X + Z) \right)^2.$$

In turns out that after applying this bound, some fortuitous simplifications occur: we end up arriving at a bound of the form

$$\text{Corr}_{\leq D} \leq \sup_f \frac{\langle c, \hat{f} \rangle}{\|M f\|} = \|c^\top M^{-1}\|$$

for a particular matrix $M$ which, crucially, is upper triangular and can thus be inverted explicitly. The resulting upper bound on $\text{Corr}_{\leq D}$ is presented in Theorem 2.2, and the full proof is given in Section 3.1. Surprisingly enough, the bounds obtained in this fashion are sharp enough to capture the conjectured recovery thresholds for problems like planted submatrix and planted dense subgraph. Some intuition for why the Jensen step (1.7) is reasonably tight is as follows: in the parameter regime where recovery is hard, the output of any low-degree polynomial of $Y$ depends almost completely on $Z$ and hardly at all on $X$, since the signal is too “small” to be seen by low-degree polynomials; thus, $f(X + Z) \approx f(Z)$ and so the inequality in (1.7) is essentially tight. We remark that applying Jensen’s inequality to both $X$ and $Z$ does not yield a useful bound on $\text{Corr}_{\leq D}$.

1.4. Notation. We use the conventions $\mathbb{N} = \{0, 1, 2, \ldots\}$ and $[N] = \{1, 2, \ldots, N\}$. For $\alpha, \beta \in \mathbb{N}^N$, define $|\alpha| = \sum_{i=1}^N \alpha_i$, $\alpha! = \prod_{i=1}^N \alpha_i!$, and (for $X \in \mathbb{R}^N$) $X^\alpha = \prod_{i=1}^N X_i^{\alpha_i}$. We use $\alpha \geq \beta$ to mean $\alpha_i \geq \beta_i$ for all $i$. The operations $\alpha + \beta$ and $\alpha - \beta$ are performed entrywise. For $\alpha, \beta \in \mathbb{N}^N$ with $\alpha \geq \beta$, define $\binom{\alpha}{\beta} = \prod_{i=1}^N \binom{\alpha_i}{\beta_i}$. The notation $\beta \leq \alpha$ means $\beta \leq \alpha$ and $\alpha \neq \beta$ (but not necessarily $\beta_i < \alpha_i$ for every $i$). On the other hand, $\beta \preceq \alpha$ simply means that $\beta \leq \alpha$ does not hold. In some cases we will restrict to $\alpha \in \{0, 1\}^N$, in which case all the same notation applies.

In some cases we will take $N = n(n+1)/2$ and view $\alpha \in \mathbb{N}^N$ as a multigraph (with self-loops allowed) on vertex set $[n]$, i.e., for each $i \leq j$, we let $\alpha_{ij}$ represent the number of edges between vertices $i$ and $j$. In this case, $V(\alpha) \subseteq [n]$ denotes the set of vertices spanned by the edges of $\alpha$.

We use standard asymptotic notation such as $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ and $o(\cdot)$, always pertaining to the limit $n \to \infty$ unless stated otherwise. Notation such as $\tilde{O}(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$ hides factors of $\log \log(n) = \log^{O(1)} n$ in the numerator or denominator. We use $f_n \ll g_n$ to mean there exists a constant $\epsilon > 0$ (not depending on $n$) such that $f_n \leq n^{-\epsilon} g_n$ for all sufficiently large $n$.

We use $1_A$ or $1[A]$ for the $\{0, 1\}$-valued indicator of an event $A$. All logarithms use the natural base unless stated otherwise.
2. Main Results. We now state our main results. We first analyze a general additive Gaussian noise model (Section 2.1) and then specialize to the planted submatrix problem (section 2.2). Then we analyze a general binary observation model (Section 2.3) and specialize to the planted dense subgraph problem (Section 2.4).

2.1. Additive Gaussian Noise Model. We consider the following general setting, which captures the planted submatrix problem but also various other popular models such as the spiked Wigner [50, 35] and (positively-spiked) Wishart [7, 8] models with any prior on the planted vector (see e.g. [87] and references therein), as well as tensor PCA [91]. A general low-degree analysis of the detection problem in the additive Gaussian noise model can be found in [74], and the above special cases are treated in [64, 63, 11, 44, 77].

Definition 2.1. In the general additive Gaussian noise model we observe
\[ Y = X + Z \]
where \( X \in \mathbb{R}^N \) is drawn from an arbitrary (but known) prior, and \( Z \) is i.i.d. \( \mathcal{N}(0,1) \), independent from \( X \). The goal is to estimate a scalar quantity \( x \in \mathbb{R} \), which is a function of \( X \).

Our main result for the additive Gaussian noise model is the following upper bound on \( \text{Corr}_{\leq D} \) as defined in (1.2) (which by Fact 1.1 implies a lower bound on \( \text{MMSE}_{\leq D} \)). The notation used here is defined in Section 1.4 above.

Theorem 2.2. In the general additive Gaussian model (Definition 2.1),
\[
\text{Corr}_{\leq D}^2 \leq \sum_{\alpha \in \mathbb{N}^N} \frac{\kappa_\alpha^2}{\alpha!},
\]
where \( \kappa_\alpha \) for \( \alpha \in \mathbb{N}^N \) is defined recursively by
\[
\kappa_\alpha = \mathbb{E}[xX^\alpha] - \sum_{0 \leq \beta \leq \alpha} \kappa_\beta \binom{\alpha}{\beta} \mathbb{E}[X^{\alpha-\beta}].
\]
The proof is given in Section 3.1; the strategy is outlined in Section 1.3. The recursive definition (2.2) requires no explicit base case; in other words, the base case is simply the case \( \alpha = 0 \) of (2.2), which is \( \kappa_0 = \mathbb{E}[x] \).

Remark 2.3. The quantity \( \kappa_\alpha \) is equal to the joint cumulant of the following collection of random variables: one instance of \( x \), and \( \alpha_i \) instances of \( X_i \) for each \( i \in [N] \). We discuss cumulants and their connection to this formula in more detail in Section 2.5.

2.2. Planted Submatrix. We now restrict our attention to a special case of the additive Gaussian model: the planted submatrix problem (which is also a variant of sparse PCA in the spiked Wigner model).

Definition 2.4. In the planted submatrix problem, we observe the \( n \times n \) matrix
\[ Y = \lambda vv^T + W \]
where \( \lambda \geq 0, v \in \{0,1\}^n \) is i.i.d. Bernoulli(\( \rho \)) for some \( \rho \in (0,1) \), and \( W \) has entries \( W_{ij} = W_{ji} \sim \mathcal{N}(0,1) \) for \( i < j \) and \( W_{ii} \sim \mathcal{N}(0,2) \), where \( \{W_{ij} : i \leq j\} \) are independent. We assume the parameters \( \lambda \) and \( \rho \) are known. The goal is to estimate \( x = v_1 \).
Prior work has extensively studied this model and variations thereof [92, 72, 9, 5, 31, 32, 78, 38, 42, 33, 60, 29, 52, 16, 19]. The statistical limits of both the detection [31] and recovery [72, 32] tasks are well-understood, as well as the computational limits of detection (for an asymmetric version of the problem) assuming the planted clique hypothesis [78, 29]. We will focus our discussion on the regime $\rho = n^{-b}$ for $b \in (0, 1)$, although the regime $\rho = \Theta(1)$ has also received attention [42, 52]. We are primarily interested in identifying the correct power and will use in our informal discussions the notation $f \ll g$ to mean there exists a constant $\epsilon > 0$ such that $f \leq n^{-\epsilon}g$ (although often only logarithmic factors will be hidden by $\ll$). We now summarize some of the relevant statistical and computational thresholds.

- **Sum test for detection**: The detection task is to hypothesis test between $Y = \lambda vv^\top + W$ and $Y = W$ with $o(1)$ error probability for both type I and type II errors. Detection is easy when $\lambda \gg (\rho \sqrt{n})^{-2}$, simply by summing all entries of $Y$ and thresholding. In Appendix B we investigate detection against different null distributions that more closely match the moments of the planted model.

- **Recovery algorithms**: When $\lambda \gg 1$, a simple entrywise thresholding algorithm can exactly recover $v$ (with probability $1-o(1)$) [72]. When $\lambda = (1 + \Omega(1))(\rho \sqrt{n})^{-1}$, the leading eigenvector of $Y$ has non-trivial correlation with $v$ (due to the “BBP transition” in random matrix theory [7, 8, 50, 35, 20]), and this can be “boosted” to exact recovery as explained in [19]. Thus, exact recovery is possible in polynomial time when $\lambda \gg \min \{1, (\rho \sqrt{n})^{-1}\}$. On the other hand, no efficient algorithm is known (even for non-trivial estimation of $v$) when $\lambda \ll \min \{1, (\rho \sqrt{n})^{-1}\}$.

- **Statistical threshold for recovery**: An estimator based on exhaustive search can achieve exact recovery when $\lambda \gg (\rho n)^{-1/2}$ [72, 32], but this is not computationally efficient. No estimator can succeed when $\lambda \ll (\rho n)^{-1/2}$ [72, 32].

The regime $\rho \gg 1/\sqrt{n}$ is of particular interest because here a detection-recovery gap appears: when $(\rho \sqrt{n})^{-2} \ll \lambda \ll (\rho \sqrt{n})^{-1}$, detection is easy but no efficient algorithm for recovery is known. Resolving the computational complexity of recovery in this regime is a problem that was left open by prior work [9, 78, 38]. Some evidence has been given that recovery is hard when $\lambda \ll (\rho \sqrt{n})^{-1}$: a reduction from planted clique shows recovery hardness for a variant of the problem, but not the canonical i.i.d. Gaussian noise model that we consider [33]; a different reduction shows recovery hardness under the assumption that a similar detection-recovery gap exists in the planted dense subgraph problem (which we discuss in Section 2.4) [29]; and a large family of MCMC methods have been shown to fail [19].

We now state our main result, which resolves the computational complexity of recovery in the low-degree framework. Part (i) shows that degree-$n^{\Omega(1)}$ polynomials fail at recovery when $\lambda \ll \min \{1, (\rho \sqrt{n})^{-1}\}$, matching the best known algorithms. Part (ii) confirms that $O(\log n)$-degree polynomials succeed in the “easy” regime $\lambda \gg \min \{1, (\rho \sqrt{n})^{-1}\}$ (provided $\rho \gg 1/n$, which ensures that the planted submatrix has non-trivial size). For context, note that the trivial estimator $f(Y) = \mathbb{E}[x] = \rho$ achieves the mean squared error $\mathbb{E}(f(Y) - x)^2 = \rho - \rho^2$. The results are non-asymptotic: they apply for all values of $n, D, \lambda, \rho$.

**Theorem 2.5.** Consider the planted submatrix problem (Definition 2.4).

(i) For any $0 < r < 1$ and $D \geq 1$, if

$$\lambda \leq \frac{r}{D(D + 1)} \min \left\{ 1, \frac{1}{\rho \sqrt{n}} \right\}$$

then

$$\text{MMSE}_{\leq D} \geq \rho - \rho^2 / (1 - r^2)^2.$$
In particular, when \( \lambda \leq \tilde{\Theta} \left( \min \left\{ 1, \frac{1}{\rho \sqrt{n}} \right\} \right) \), the MMSE \( \leq \text{polylog}(n) \) is at most \( o(\rho^2) \) smaller than the error of the trivial estimator.

(ii) For any \( 0 < r < 1 \) and odd \( D \geq 1 \), if

\[
\lambda \geq \frac{24}{r} \sqrt{\log 8 + 2D \log(9/\rho)} \min \left\{ 1, \frac{1}{\rho \sqrt{n}} \right\}
\]

and

\[
\frac{324}{r^2 n} \left( \log 8 + 2D \log(9/\rho) \right) \leq \rho \leq 1/2
\]

then

\[
\text{MMSE}_{\leq D} \leq D^2 r D^{-1}.
\]

In particular, provided that \( \tilde{\Theta} \left( \frac{1}{n} \right) \leq \rho \leq \frac{1}{2} \) and \( \lambda \geq \tilde{\Theta} \left( \min \left\{ 1, \frac{1}{\rho \sqrt{n}} \right\} \right) \), the MMSE \( \leq \log n \) is at most \( n^{\log \log n} \).

The proof of part (i) is given in Section 3.2, while part (ii) follows by combining Theorems 4.2 and 4.3 from Section 4. In Appendix E we give a more refined analysis that suggests a sharp computational threshold at \( \lambda = (\rho \sqrt{n})^{-1} \) in the regime \( 1/\sqrt{n} \ll \rho \ll 1 \), although these results are limited to fairly low degree: \( D \leq \log_2(1/\rho) - 1 \). This sharp constant in the threshold matches the one discovered in [60] (see also [43]), where an AMP-style algorithm is shown to succeed above the threshold.

2.2.1. Discussion and future directions. While we have pinned down the correct power of \( n \) for the critical value of \( \lambda \), some more fine-grained questions remain open. The first is to extend the result to larger values of \( D \). It has been conjectured [63] that for a broad class of high-dimensional problems, degree-\( D \) polynomials are as powerful as \( n^{\tilde{O}(D)} \)-time algorithms, where \( \tilde{O}(\cdot) \) hides factors of \( \log n \). This conjecture is consistent with the best known algorithms for tensor PCA and sparse PCA: see [74, 44]. In the setting of planted submatrix, the best known algorithms in the “hard” regime \( \lambda \ll \min \{ 1, (\rho \sqrt{n})^{-1} \} \) run in time \( n^{\tilde{O}(\lambda^{-2})} \) [44, 62], and a large class of MCMC methods can do no better [19]. To give low-degree evidence that this is the best possible runtime, one could hope to prove that MMSE \( \leq D \) is large for all \( D \ll \lambda^{-2} \) whenever \( \lambda \ll \min \{ 1, (\rho \sqrt{n})^{-1} \} \).

An additional direction for future work is to more precisely understand the boundary between the “easy” and “hard” regimes. We have done this for very low degree \( D \leq \log_2(1/\rho) - 1 \) in Appendix E, and ideally this would be extended to some \( D = \omega(\log n) \).

2.3. Binary Observation Model. We next consider a different setting, where the observed variables are binary-valued. This captures, for instance, various problems where the observation is a random graph.

**Definition 2.6.** The general binary observation model is defined as follows. First, a signal \( X \in [0, 1]^N \) is drawn from an arbitrary (but known) prior. We observe \( Y \in \{0, 1\}^N \) where \( E[Y_i | X] = X_i \) and \( \{Y_i\} \) are conditionally independent given \( X \). Assume that the law of \( X \) is supported on \([\tau_0, \tau_1]^N \) where \( 0 < \tau_0 \leq \tau_1 < 1 \). The goal is to estimate a scalar quantity \( x \in \mathbb{R} \), which is a function of \( X \).

In this setting, we obtain an upper bound on \( \operatorname{Corr}_{\leq D} \) that is similar to the additive Gaussian setting.
**Theorem 2.7.** In the general binary observation model (Definition 2.6),

\[ \text{Corr}_{\leq D}^2 \leq \sum_{\alpha \in \{0,1\}^N} \frac{\kappa_\alpha^2}{\tau_0(1 - \tau_1)^{|\alpha|}}, \]

where \( \kappa_\alpha \) for \( \alpha \in \{0,1\}^N \) is defined recursively by

\[ \kappa_\alpha = \mathbb{E}[x X^\alpha] - \sum_{0 \leq \beta \leq \alpha} \kappa_\beta \mathbb{E}[X^{\alpha - \beta}]. \]

The proof is similar to that of Theorem 2.2 and is deferred to Appendix G.1. Note that this is the same definition of \( \kappa_\alpha \) as in Theorem 2.2; the factor of \( \binom{|\alpha|}{2} \) is not needed here since we are restricting to \( \alpha, \beta \in \{0,1\}^N \). In particular, \( \kappa_\alpha \) can be interpreted as a certain joint cumulant (see Claim 2.14).

**2.4. Planted Dense Subgraph.** We now specialize the result of the previous section to the planted dense subgraph problem, which can be thought of as the analogue of the planted submatrix problem in random graphs.

**Definition 2.8.** In the planted dense subgraph problem, we observe a random graph \( Y \in \{0,1\}^{(n)} \) generated as follows. First a planted signal \( v \in \{0,1\}^n \) is drawn with i.i.d. Bernoulli(\( \rho \)) entries. Conditioned on \( v \), the value \( Y_{ij} \sim \text{Bernoulli}(q_0 + (q_1 - q_0)v_i v_j) \) is sampled independently for each \( i < j \), for some \( q_0, q_1 \in [0,1] \). We assume that the parameters \( \rho, q_0, q_1 \) are known. The goal is to estimate \( x = v_1 \).

Here, \( q_1 \) represents the connection probability within the planted subgraph, and \( q_0 \) represents the connection probability elsewhere. We will assume \( q_0 \leq q_1 \) without loss of generality (since otherwise one can consider the complement graph instead).

Like the planted submatrix problem, this problem (and variants) have been extensively studied [22, 6, 3, 94, 38, 59, 39]. The statistical limits of detection [6, 94, 59] and recovery [3, 38] are known, as well as the computational limits of detection (assuming the planted clique hypothesis) [59]. Resolving the computational limits of recovery is stated as an open problem in [38, 59].

Since our bound on \( \text{Corr}_{\leq D} \) for the binary model (Theorem 2.7) is similar to our bound in the Gaussian model (Theorem 2.2), we are able to export our results on planted submatrix to this setting and obtain the following.

**Theorem 2.9.** Consider the planted dense subgraph problem (Definition 2.8) with \( \rho \leq 1/2 \) and \( q_0 \leq q_1 \).

(i) For any \( 0 < r < 1 \) and \( D \geq 1 \), if

\[ \frac{q_1 - q_0}{\sqrt{q_0(1 - q_1)}} \leq \frac{r}{D(D + 1)} \min \left\{ 1, \frac{1}{\rho \sqrt{n}} \right\} \]

then

\[ \text{MMSE}_{\leq D} \geq \rho - \rho^2/(1 - r^2)^2. \]

(ii) Let

\[ \nu := \min \{ \rho, q_0, 1 - q_1 \}. \]
For any $0 < r < 1$ and odd $D \geq 1$, if
\[
\frac{(q_1 - q_0)^2}{q_0} \geq \frac{216}{r^2 \rho^2 (n-1)} \left[ \log 4 + 3D \log \left( \frac{9}{\nu} \right) \right]
\]
and
\[
q_1 \rho \geq \frac{864}{r^2 (n-1)} \left[ \log 4 + 3D \log \left( \frac{9}{\nu} \right) \right]
\]
then
\[
\text{MMSE}_{\leq D} \leq D^2 r^{-D-1}.
\]

The proof of part (i) is given in Appendix G.2, while part (ii) is an immediate consequence of Theorem I.5 in Appendix I. For context, the trivial estimator $f(Y) = E[x] = \rho$ achieves the mean squared error $E((f(Y) - x)^2) = \rho - \rho^2$. Thus, part (i) states that degree-$n^{-\Omega(1)}$ polynomials fail at recovery when
\[
\frac{q_1 - q_0}{\sqrt{q_0(1-q_1)}} \ll \min \left\{ 1, \frac{1}{\rho \sqrt{n}} \right\},
\]
while part (ii) states that (under the mild condition $\log(1/\nu) = n^{o(1)}$) $O(\log n)$-degree polynomials succeed when
\[
\frac{q_1 - q_0}{\sqrt{q_0}} \gg \frac{1}{\rho \sqrt{n}} \quad \text{and} \quad q_1 \rho \gg \frac{1}{n}.
\]

We discuss the implications of these bounds below.

2.4.1. Discussion and future directions. Theorem 2.9 (summarized in (2.5) and (2.6)) resolves (in the low-degree framework) the computational complexity of recovery in a number of previously-studied settings. For instance, [59] considers the regime $q_1 = cq_0 = \Theta(n^{-a})$ and $\rho = \Theta(n^{b-1})$ for constants $a \in (0,2)$, $b \in (0,1)$, and $c > 1$. In this setting, Theorem 2.9(i) shows low-degree hardness of recovery at degree $n^{\Omega(1)}$ whenever $b < (1+a)/2$. This is precisely the regime where planted-clique-hardness of recovery was left open by [59] (see Figure 2 of [59]). Furthermore, Theorem 2.9(ii) gives a matching upper bound, showing that low-degree polynomials succeed whenever $b > (1+a)/2$; this coincides with the best known polynomial-time algorithms (see [59]).

A more general regime of parameters is considered by [29] (for a close variant of the problem where $v$ has exactly $\rho m$ nonzero entries). They state the PDS (planted dense subgraph) recovery conjecture which postulates hardness of exact recovery under the conditions
\[
\liminf_{n \to \infty} \log_n \rho > -1/2 \quad \text{and} \quad \limsup_{n \to \infty} \log_n \left( \frac{\rho^2 (q_1 - q_0)^2}{q_0 (1-q_0)} \right) < -1,
\]
and they show that this conjecture implies (via average-case reductions) detection-recovery gaps in some other problems: biased sparse PCA and biclustering (which is an asymmetric version of planted submatrix). Prior to our work, concrete evidence for the PDS recovery conjecture was somewhat lacking (see the discussion following Conjecture 2.8 of [38]). Our results establish low-degree hardness of recovery in much (but not all) of the conjectured hard regime; for instance, if $q_0 \leq q_1 \leq 1 - \Omega(1)$ (or even if $q_0 \leq q_1 \leq 1 - n^{-o(1)}$) then Theorem 2.9(i) shows low-degree hardness at degree $n^{\Omega(1)}$ whenever (2.7) holds.

One special case not covered by Theorem 2.9 is the planted clique problem, where $q_1 = 1$ (since the denominator of the left-hand side of (2.5) is zero in this case). While no detection-recovery gap is expected here, it is still non-trivial to bound the low-degree MMSE. In Appendix H we show that our techniques can be extended to the planted clique problem, giving
matching upper and lower bounds for recovery by low-degree polynomials; these coincide with the widely-conjectured computational threshold for planted clique. The proof of the lower bound requires a modification of the argument used to prove Theorem 2.7; while Theorem 2.9(i) can be proved from Theorem 2.7 with relative ease by leveraging our existing calculations for planted submatrix, the planted clique result requires a specialized treatment “from scratch”.

We note that the upper and lower bounds in Theorem 2.9 (summarized in (2.5) and (2.6)) match in many parameter regimes (such as the one considered by [59], as discussed above), but not all. The condition \( q_1 \rho \gg 1/n \) in the upper bound is natural because it ensures that each vertex in the planted submatrix has at least one neighbor in the planted submatrix; without this, recovery is impossible. Assuming \( q_1 \rho \gg 1/n \), the upper and lower bounds match so long as \( q_1 \) is not extremely close to 1 (specifically \( q_1 \leq 1 - n^{-o(1)} \)) and \( \rho \gg 1/\sqrt{n} \). On the other hand, a more elusive regime is \( \rho \ll 1/\sqrt{n} \) with \( q_1 = \Theta(n^{-a}) \) and \( q_0 = \Theta(n^{-b}) \) for positive constants \( a < b \); we do not expect a detection-recovery gap here, but even the detection question remains largely unanswered by prior work. The best known polynomial-time algorithm for this regime seems to be that of [22] (they give a proof sketch in Section 3.2), which is based on counting caterpillar graphs and succeeds when \( a < cb \) where \( \rho = \Theta(n^{-c}) \). We suspect that it may be possible to give a better algorithm that succeeds in the regime \( a < b/2 \), matching our lower bound. However, this might require a rather involved subgraph-counting argument as in [14], and we leave this for future work.

2.5. Cumulants. Here we give a brief overview of cumulants and some of their basic properties. As we show in Claim 2.14 below, the quantities \( \kappa_\alpha \) that appear in our main results (Theorems 2.2 and 2.7) can be interpreted as certain joint cumulants. Later on, this connection will be convenient for deducing certain helpful properties of \( \kappa_\alpha \). For more details on cumulants, we refer the reader to e.g. [85].

**Definition 2.10.** Let \( X_1, \ldots, X_n \) be jointly-distributed random variables. Their cumulative generating function is the function

\[
K(t_1, \ldots, t_n) = \log \mathbb{E} \left[ \exp \left( \sum_{i=1}^n t_i X_i \right) \right],
\]

and their joint cumulant is the quantity

\[
\kappa(X_1, \ldots, X_n) = \left. \left( \prod_{i=1}^n \frac{\partial}{\partial t_i} K(t_1, \ldots, t_n) \right) \right|_{t_1 = \cdots = t_n = 0}
\]

\[
= \sum_{\pi \in \mathcal{P}} (|\pi| - 1)! (-1)^{|\pi| - 1} \prod_{B \in b(\pi)} \mathbb{E} \left[ \prod_{i \in B} X_i \right],
\]

where \( \mathcal{P} \) is the set of all partitions of \([n]\) (here we are considering partitions of \( n \) labeled elements into nonempty unlabeled parts), and for a given \( \pi \in \mathcal{P} \) we use \( b(\pi) \) to denote the set of all parts of the partition and \( |\pi| \) to denote the number of parts in the partition.

Cumulants are an alternative to moments in specifying a probability distribution. The joint cumulant \( \kappa(\cdots) \) is symmetric, i.e., it depends only on the (multi)set of random variables \( \{X_1, \ldots, X_n\} \) and not the order in which they are listed. The cumulants enjoy some convenient properties, which we record here. The proofs are given in Appendix D for completeness.
PROPOSITION 2.11. If \( a, b \geq 1 \) and \( X_1, X_2, \ldots, X_a, Y_1, \ldots, Y_b \) are random variables with \( \{ Y_i \}_{i \in [a]} \) independent from \( \{ X_j \}_{j \in [b]} \), then
\[
\kappa(X_1, X_2, \ldots, X_a, Y_1, \ldots, Y_b) = 0.
\]

PROPOSITION 2.12. If \( X_1, \ldots, X_n \) and \( Y_1, \ldots, Y_n \) are random variables with \( \{ X_i \}_{i \in [n]} \) independent from \( \{ Y_i \}_{i \in [n]} \), then
\[
\kappa(X_1 + Y_1, \ldots, X_n + Y_n) = \kappa(X_1, \ldots, X_n) + \kappa(Y_1, \ldots, Y_n).
\]

PROPOSITION 2.13. The joint cumulant is invariant under constant shifts and is scaled by constant multiplication. That is, if \( X_1, \ldots, X_n \) are jointly-distributed random variables and \( c \) is any constant, then
\[
\kappa(X_1 + c, X_2, \ldots, X_n) = \kappa(X_1, \ldots, X_n) + c \cdot \mathbb{1}[n = 1],
\]
and
\[
\kappa(cX_1, X_2, \ldots, X_n) = c \cdot \kappa(X_1, \ldots, X_n).
\]

In equations (2.2) and (2.4), we introduced quantities \( \kappa_\alpha \) for \( \alpha \in \mathbb{N}^N \). Here we will think of \( \alpha \) as a multiset \( \{ a_1, \ldots, a_m \} \) which contains \( a_i \) copies of \( i \) for all \( i \in [N] \). We can show that these are the joint cumulants of a (multi)set of entries of the planted signal.

CLAIM 2.14. If \( \alpha = \{ a_1, \ldots, a_m \} \), then
\[
\kappa_\alpha = \kappa(x, X_{a_1}, \ldots, X_{a_m}).
\]

We defer the proof to Appendix D. We remark that we are not aware of a “deeper” reason for why cumulants appear in our context. One particularly important consequence of Claim 2.14 is that when \( x \) is independent of \( X_{a_1}, \ldots, X_{a_m} \), we have \( \kappa_\alpha = 0 \); this is crucial for us in showing that (2.1) and (2.3) are bounded, since many of the terms are zero. This is in contrast to what happens in the case of detection, where terms corresponding to such \( \alpha \) do contribute to make \( \| L^{\leq D} \| \) large.

3. Lower Bounds. In this section, we prove our lower bounds on \( \text{MMSE}_{\leq D} \) (or equivalently, upper bounds on \( \text{Corr}_{\leq D} \)) for the additive Gaussian model (Theorem 2.2) and planted submatrix problem (Theorem 2.5(i)). The analogous proofs for binary-valued problems, including planted dense subgraph, are deferred to Appendix G.

3.1. Additive Gaussian Noise Model. In this section we prove Theorem 2.2. We will work with the Hermite polynomials, which are orthogonal polynomials with respect to Gaussian measure (see [93] for a standard reference). Recall the notation regarding \( \alpha \in \mathbb{N}^N \), defined in Section 1.4. Let \( (h_k)_{k \in \mathbb{N}} \) be the normalized Hermite polynomials \( h_k = \frac{1}{\sqrt{k!}} H_k \) where \( (H_k)_{k \in \mathbb{N}} \) are defined by the recurrence
\[
H_0(z) = 1, \quad H_1(z) = z, \quad H_{k+1}(z) = zH_k(z) - kH_{k-1}(z) \quad \text{for} \quad k \geq 1.
\]
The normalization ensures orthonormality \( \mathbb{E}_{z \sim \mathcal{N}(0,1)}[h_k(z) \overline{h}_l(z)] = \mathbb{1}_{k=l} \). In particular, \( \mathbb{E}_{z \sim \mathcal{N}(0,1)}[h_k(z)] = 0 \) for all \( k \geq 1 \). Also, for \( \alpha \in \mathbb{N}^N \) and \( Z \in \mathbb{R}^N \), let \( h_\alpha(Z) = \prod_{i \in [N]} h_{\alpha_i}(Z_i) \); these form an orthonormal basis with respect to \( \mathcal{N}(0,1)^{\otimes N} \), i.e., if \( Z \) has i.i.d. \( \mathcal{N}(0,1) \) entries then \( \mathbb{E}[h_\alpha(Z) \overline{h}_\beta(Z)] = \mathbb{1}_{\alpha = \beta} \). A shifted Hermite polynomial can be expanded in the Hermite basis as follows.
PROPOSITION 3.1. For any \( k \in \mathbb{N} \) and \( z, \mu \in \mathbb{R} \),

\[
(3.2) \quad H_k(z + \mu) = \sum_{\ell=0}^{k} \frac{k!}{\ell!} \mu^{k-\ell} H_\ell(z)
\]

and so

\[
(3.3) \quad h_k(z + \mu) = \sum_{\ell=0}^{k} \sqrt{\frac{\ell!}{k!}} \frac{k!}{\ell!} \mu^{k-\ell} h_\ell(z).
\]

In particular,

\[
(3.4) \quad \mathbb{E}_{z \sim \mathcal{N}(0,1)} h_k(z + \mu) = \frac{\mu^k}{\sqrt{k!}}.
\]

These facts are well known (see e.g. page 254 of [79]), but we provide a proof in Appendix F.1 for completeness.

PROOF OF THEOREM 2.2. For a degree-\( D \) polynomial \( f \), write the Hermite expansion

\[
f(Y) = \sum_{|\alpha| \leq D} \hat{f}_\alpha h_\alpha(Y).
\]

Using (3.4), we have

\[
\mathbb{E}[f(Y)^2] \geq \mathbb{E}_X \left( \mathbb{E}_Z f(X + Z) \right)^2 =: \mathbb{E}_Z g(Z)^2 = \|\tilde{g}\|^2,
\]

where \( \{\hat{g}_\alpha\} \) are the Hermite coefficients of the function \( g(Z) = \mathbb{E}_X f(X + Z) \), and \( \|\tilde{g}\|^2 \) denotes \( \sum_{|\alpha| \leq D} \hat{g}_\alpha^2 \). We may calculate the \( \{\hat{g}_\alpha\} \) directly, as a function of the \( \{\hat{f}_\alpha\} \):

\[
g(Z) = \mathbb{E}_X f(X + Z) = \sum_{\alpha} \hat{f}_\alpha \mathbb{E}_X h_\alpha(X + Z)
\]

\[
= \sum_{\alpha} \hat{f}_\alpha \mathbb{E}_X \prod_i h_{\alpha_i}(X_i + Z_i) \quad \text{(by definition of } h_\alpha)\]

\[
= \sum_{\alpha} \hat{f}_\alpha \mathbb{E}_X \prod_i \sum_{\ell=0}^{\alpha_i} \sqrt{\frac{\ell!}{\alpha_i!}} \frac{\alpha_i!}{\ell!} X_i^{\alpha_i-\ell} h_\ell(Z_i) \quad \text{(by Proposition 3.1)}
\]

\[
= \sum_{\alpha} \hat{f}_\alpha \mathbb{E}_X \sum_{0 \leq \beta \leq \alpha} h_\beta(Z) \mathbb{E}_X [X^{\alpha-\beta}] \sqrt{\frac{\beta!}{\alpha!}} \frac{\alpha!}{\beta!} X_i^{\alpha_i-\beta} h_\beta(Z_i)
\]

\[
= \sum_{\alpha} \hat{f}_\alpha \sum_{0 \leq \beta \leq \alpha} h_\beta(Z) \mathbb{E}_X [X^{\alpha-\beta}] \sqrt{\frac{\beta!}{\alpha!}} \frac{\alpha!}{\beta!} \frac{\alpha_i!}{\beta_i!} X_i^{\alpha_i-\beta} h_\beta(Z_i)
\]

To summarize, \( \mathbb{E}[f(Y)^2] = \langle c, \hat{f} \rangle \) where \( c = (c_\alpha)_{|\alpha| \leq D} \) is defined by

\[
c_\alpha := \mathbb{E}_X [X^{\alpha}] / \sqrt{\alpha!},
\]

and \( \mathbb{E}[f(Y)^2] \geq \|\tilde{g}\|^2 \) where \( \tilde{g} = M \hat{f} \) with

\[
M_{\beta \alpha} := \mathbb{E}_X [X^{\alpha-\beta}] \sqrt{\frac{\beta!}{\alpha!}} \frac{\alpha!}{\beta!} \frac{\alpha_i!}{\beta_i!} X_i^{\alpha_i-\beta} h_\beta(Z_i).
\]
for all $\alpha, \beta \in \mathbb{N}^N$ with $|\alpha| \leq D, |\beta| \leq D$. Note that $M$ is upper triangular with 1’s on the diagonal, and is thus invertible. Now

$$\text{Corr}_{\leq D} \leq \sup_{\hat{f} \neq 0} \frac{\langle c, \hat{f} \rangle}{\|M \hat{f}\|} = \sup_{\hat{g} \neq 0} \frac{c^T M^{-1} \hat{g}}{\|\hat{g}\|} = \|c^T M^{-1}\| =: \|w\|$$

where $w^T M = c^T$ (since the optimizer for $\hat{g}$ is $(M^{-1})^T c$). Solve for $w$ recursively:

$$c_\alpha = \sum_\beta w_\beta M_{\beta \alpha} = \sum_{\beta \leq \alpha} w_\beta \mathbb{E}[X^{\alpha - \beta}] \sqrt{\frac{\beta!}{\alpha!}} \binom{\alpha}{\beta}$$

and so

$$w_\alpha = c_\alpha - \sum_{\beta \leq \alpha} w_\beta \mathbb{E}[X^{\alpha - \beta}] \sqrt{\frac{\beta!}{\alpha!}} \binom{\alpha}{\beta}.$$

Equivalently, $w_\alpha = \kappa_\alpha / \sqrt{\alpha!}$ where

$$\kappa_\alpha = \mathbb{E}[x X^\alpha] - \sum_{\beta \leq \alpha} \kappa_\beta \binom{\alpha}{\beta} \mathbb{E}[X^{\alpha - \beta}].$$

We have shown

$$\text{Corr}_{\leq D}^2 \leq \|w\|^2 = \sum_{|\alpha| \leq D} w^2_\alpha = \sum_{|\alpha| \leq D} \frac{\kappa^2_\alpha}{\alpha!},$$

completing the proof. \qed

3.2. Planted Submatrix. In this section we prove Theorem 2.5(i) using Theorem 2.2. To cast planted submatrix as a special case of the additive Gaussian noise model, we take $X = (X_{ij})_{i \leq j}$ defined by $X_{ij} = \lambda v_i v_j$. Note that we have removed the redundant lower-triangular part of the matrix, and decreased the noise on the diagonal from $\mathcal{N}(0, 2)$ to $\mathcal{N}(0, 1)$. This modification on the diagonal can only increase $\text{Corr}_{\leq D}$ (see Claim A.2 in Appendix A), and so an upper bound on $\text{Corr}_{\leq D}$ in this new model implies the same upper bound on $\text{Corr}_{\leq D}$ in the original model.

We will think of $\alpha = (\alpha_{ij})_{i \leq j}$ (where $\alpha_{ij} \in \mathbb{N}$) as a multigraph (with self-loops allowed) on vertex set $[n]$, where $\alpha_{ij}$ represents the number of edges between vertices $i$ and $j$.

**Lemma 3.2.** If $\alpha$ has a nonempty connected component that does not contain vertex 1, then $\kappa_\alpha = 0$. (In particular, $\kappa_\alpha = 0$ whenever $\alpha$ is disconnected.)

Here, nonempty means that the connected component contains at least one edge.

**Remark 3.3.** Lemma 3.2 is crucial to separating recovery from detection, as it allows us to restrict our attention to connected multigraphs. As illustrated by the proof of Proposition C.1, the multigraphs responsible for making detection easy are highly disconnected.

Using the cumulant interpretation of $\kappa_\alpha$ (Remark 2.3), Lemma 3.2 follows easily from the following basic property of cumulants: the joint cumulant of a collection of random variables is zero whenever the random variables can be partitioned into two nonempty parts that are independent from each other; see Section 2.5 and particularly Proposition 2.11. We also give an alternative self-contained proof of Lemma 3.2 in Appendix F.2.
PROOF OF LEMMA 3.2. From Claim 2.14, \( \kappa_\alpha \) is the joint cumulant of \( v_1 \) along with the edge variables \( X_{ij} = \lambda v_i v_j \) for each edge \((i, j)\) of \( \alpha \). Since the \( v_i \) are sampled independently, if there is a connected component \( C \) not containing \( v_1 \), then \( v_1 \) and the edge variables in \( C \) are independent from the edge variables in \( C \). This gives \( \kappa_\alpha = 0 \) by Proposition 2.11.

Let \( V(\alpha) \subseteq [n] \) denote the set of vertices spanned by \( \alpha \).

**Lemma 3.4.** \( \kappa_0 = \rho \) and for \( |\alpha| \geq 1 \),

\[
|\kappa_\alpha| \leq (|\alpha| + 1)^{|\alpha|} |v_1| |E(\gamma)|. 
\]

**Proof.** Proceed by induction on \( |\alpha| \). In the base case \( |\alpha| = 0 \), we have \( \kappa_0 = \rho \). For \( |\alpha| \geq 1 \), by Lemma 3.2 we can assume \( \alpha \) is connected and spans vertex 1. We have by the triangle inequality,

\[
|\kappa_\alpha| \leq |E[xX^\alpha]| + \sum_{\beta \subseteq \alpha} |\kappa_\beta| \left( \begin{array}{c} \alpha \\ \beta \end{array} \right) |E[X^{\alpha-\beta}]|.
\]

Now, for any multigraph \( \gamma \), \( E[X^\gamma] = E[\prod_{(i,j) \in E(\gamma)} \lambda v_i v_j] = \lambda^{\gamma} \cdot E[\prod_{i \in V(\gamma)} v_i^{\deg(i)}] = \lambda^{\gamma} \rho^{\gamma(|\gamma|)} \), since the \( v_i \) are independent Bernoulli(\( \rho \)). Similarly, \( E[xX^\gamma] = \lambda^{\gamma} \rho^{\gamma(|\gamma|)+1} \).

Thus, we may bound the above:

\[
\leq \lambda^{|\alpha|} \rho^{V(\alpha)} + \sum_{\beta \subseteq \alpha} |\kappa_\beta| \left( \begin{array}{c} \alpha \\ \beta \end{array} \right) \lambda^{|\alpha-\beta|} \rho^{V(\alpha-\beta)}
\]

\[
= \lambda^{|\alpha|} \rho^{V(\alpha)} + \lambda^{|\alpha|} \rho^{1+|V(\alpha)|} + \sum_{0 \neq \beta \subseteq \alpha} |\kappa_\beta| \left( \begin{array}{c} \alpha \\ \beta \end{array} \right) \lambda^{|\alpha-\beta|} \rho^{V(\alpha-\beta)},
\]

and now applying the induction hypothesis to \( \kappa_\beta \) for all \( 0 \neq \beta \not\subseteq \alpha \),

\[
\leq 2\lambda^{|\alpha|} \rho^{V(\alpha)} + \sum_{0 \neq \beta \subseteq \alpha} (|\beta| + 1)^{|\beta|} \lambda^{|\beta|} \rho^{V(\beta)} \left( \begin{array}{c} \alpha \\ \beta \end{array} \right) \lambda^{|\alpha-\beta|} \rho^{V(\alpha-\beta)}.
\]

Since \( |V(\beta)| + |V(\alpha-\beta)| \geq |V(\alpha)| \) and \( \rho \leq 1 \), we may pull out a factor of \( \rho^{V(\alpha)} \),

\[
\leq \lambda^{|\alpha|} \rho^{V(\alpha)} \left[ 2 + \sum_{0 \neq \beta \subseteq \alpha} (|\beta| + 1)^{|\beta|} \left( \begin{array}{c} \alpha \\ \beta \end{array} \right) \right],
\]

and now, we bound the parenthesized quantity in a straightforward manner:

\[
= \lambda^{|\alpha|} \rho^{V(\alpha)} \left[ 2 + \sum_{\ell=1}^{|\alpha|-1} (\ell + 1)^\ell \binom{|\alpha|}{\ell} \right]
\]

\[
\leq \lambda^{|\alpha|} \rho^{V(\alpha)} \left[ 2 + \sum_{\ell=1}^{|\alpha|-1} |\alpha|^{\ell} \binom{|\alpha|}{\ell} \right]
\]

\[
\leq \lambda^{|\alpha|} \rho^{V(\alpha)} \sum_{\ell=0}^{|\alpha|} |\alpha|^{\ell} \binom{|\alpha|}{\ell} \leq \lambda^{|\alpha|} \rho^{V(\alpha)} (|\alpha| + 1)^{|\alpha|},
\]
where the last step used the binomial theorem. This completes the proof.

We will combine our bounds on $\kappa_\alpha$ with a bound on the number of multigraphs $\alpha$ that we must consider.

**Lemma 3.5.** For integers $d \geq 1$ and $0 \leq h \leq d$, the number of connected multigraphs $\alpha$ on vertex set $[n]$ such that (i) $|\alpha| = d$, (ii) $1 \in V(\alpha)$, and (iii) $|V(\alpha)| = d + 1 - h$, is at most $(dn)^d \left( \frac{d}{n} \right)^h$.

**Proof.** For any such $\alpha$, we can order the edges so that every prefix of edges is connected, the first edge spans vertex 1, each of the first $d - h$ edges spans a new vertex (not including vertex 1), and the last $h$ edges do not span any new vertices. For the first $d - h$ steps there are $\leq (dn)^d$ choices, and in the last $h$ steps there are $\leq d^2$ choices. In total this gives $(dn)^{d-h}(d^2)^h = (dn)^d(d/n)^h$.

**Proof of Theorem 2.5(i).** Using Theorem 2.2 and Lemma 3.2, we have

$$\Corr_{\leq D}^2 \leq \sum_{0 \leq |\alpha| \leq D} \frac{\kappa_\alpha^2}{\alpha!} \leq \rho^2 + \sum_{1 \leq |\alpha| \leq D} \kappa_\alpha^2_{\alpha_{\alpha \neq V(\alpha), \alpha \text{ connected}}}.$$

Now, splitting the sum over $\alpha$ according to the number of edges $d$ and the number of vertices $d + 1 - h$ (there are at most $d + 1$ vertices, as $\alpha$ is connected), and applying our bounds on the magnitude of the corresponding $\kappa_\alpha$ and on the number of such $\alpha$ from Lemmas 3.4 and 3.5,

$$\leq \rho^2 + \sum_{d=1}^D \sum_{h=0}^d (dn)^d \left( \frac{d}{n} \right)^h \left[ (d + 1)^d \chi^d \rho^{d+1-h} \right]^2 \leq \rho^2 \cdot \sum_{d=0}^D \sum_{h=0}^d \left[ d(d + 1)^2 \lambda^2 \rho^2 n \right]^d \left( \frac{d}{\rho^2 n} \right)^h \leq \rho^2 \sum_{d=0}^D \sum_{h=0}^d \left[ D(D + 1)^2 \lambda^2 \rho^2 n \right]^d \left( \frac{D}{\rho^2 n} \right)^h \leq \rho^2 \sum_{d=0}^D \sum_{h=0}^d \left[ D^2(D + 1)^2 \lambda^2 \right]^h \sum_{d=h}^D \left[ D(D + 1)^2 \lambda^2 \rho^2 n \right]^{d-h} \leq \rho^2 \sum_{h=0}^D \sum_{d=h}^D \sum_{d'}^{d-h} \rho^2 = \rho^2 \left( \frac{1 - r^2}{1 - 2 \rho^2(1 - r^2)} \right).$$

The result now follows from Fact 1.1.

**4. Upper Bounds.** In this section we present our upper bounds on $\text{MMSE}_{\leq D}$ for the planted submatrix problem (see Appendix I for analogous results on planted dense subgraph). In Theorems 4.2 and 4.3 below, we will show that two standard algorithms (diagonal thresholding and power iteration, respectively) can be implemented using low-degree polynomials. Combining these two theorems immediately yields the proof of Theorem 2.5(ii). We remark that Theorems 4.2 and 4.3 are somewhat unsurprising, yet are important to establish in order to make our matching lower bounds meaningful.

We start with a useful subroutine: a polynomial approximation to the threshold function.
Proposition 4.1. For any integer $k \geq 0$, there is a degree-$2k+1$ polynomial $\tau = \tau_k : \mathbb{R} \to \mathbb{R}$ such that for any $\ell \in \{0, 1\}$ and any $0 \leq \Delta \leq 1/2$, 
\[ |\tau(y) - \ell| \leq (k + 1/2)(6\Delta)^k \]
whenever $|y - \ell| \leq \Delta$. 

The proof is deferred to Appendix I.1.

4.1. Diagonal Thresholding. The simple diagonal thresholding algorithm simply picks out the largest diagonal entries of the input matrix $[70, 4, 72]$. Here we show how to construct a low-degree polynomial based on this idea that achieves small mean squared error when $\lambda \gg 1$.

Theorem 4.2. Consider the planted submatrix problem (Definition 2.4) with $\rho \leq 1/2$, and let $\tau_k$ be as in Proposition 4.1. Let $k \geq 0$ and consider the polynomial $f(Y) = \tau_k(Y_{11}/\lambda) = \tau_k(v_1 + W_{11}/\lambda)$ of degree $D := 2k + 1$. For any $0 < r < 1$, if
\[
\lambda \geq \frac{12}{r} \sqrt{\log 4 + 2D\log(9/\rho)}
\]
then
\[ \mathbb{E}(f(Y) - x)^2 \leq D^2r^{D-1}. \]

The proof is deferred to Appendix I.

4.2. Power Iteration. When $\lambda \geq (1 + \Omega(1))(\rho\sqrt{n})^{-1}$, it is well-known in random matrix theory that the leading eigenvector of $Y$ is correlated with $v$ [50, 35, 20]. Furthermore, the associated eigenvalue is larger than the other eigenvalues by a constant factor, and so the leading eigenvector can be well-approximated by $O(\log n)$ rounds of power iteration.

To simplify our analysis, we will consider a single round of power iteration starting from the all-ones vector, followed by thresholding. While this does not capture the sharp threshold above, we show that this achieves small mean squared error provided $\lambda \gg (\rho\sqrt{n})^{-1}$ and $\rho \gg 1/n$.

Theorem 4.3. Consider the planted submatrix problem (Definition 2.4) with $\rho \leq 1/2$, and let $\tau_k$ be as in Proposition 4.1. Let $k \geq 0$ and consider the polynomial
\[
f(Y) = \tau_k \left( \frac{1}{\lambda \rho n} \sum_{i=1}^{n} Y_{1i} \right) = \tau_k \left( \frac{1}{\lambda \rho n} \sum_{i=1}^{n} (\lambda v_1 v_i + W_{1i}) \right)
\]
of degree $D := 2k + 1$. For any $0 < r < 1$, if
\begin{align}
\lambda &\geq \frac{24}{r \rho \sqrt{n}} \sqrt{\log 8 + 2D\log(9/\rho)} \\
\rho &\geq \frac{324}{r^2 n} [\log 8 + 2D\log(9/\rho)]
\end{align}
then
\[ \mathbb{E}(f(Y) - x)^2 \leq D^2r^{D-1}. \]

The proof is deferred to Appendix I.
Acknowledgements. For helpful discussions, we are grateful to Afonso Bandeira, Matthew Brennan, Jingqiu Ding, David Gamarnik, Sam Hopkins, Frederic Koehler, Tim Kunisky, Jerry Li, Jonathan Niles-Weed, and Ilias Zadik. We thank the anonymous reviewers for their helpful comments. This work was done while TS was virtually visiting the Microsoft Research Machine Learning and Optimization group. ASW was partially supported by NSF grant DMS-1712730 and by the Simons Collaboration on Algorithms and Geometry.

SUPPLEMENTARY MATERIAL

Supplement to “Computational Barriers to Estimation from Low-Degree Polynomials”. Contains technical appendices A through J.

REFERENCES


