Tractable Gaussian Phase Retrieval with Heavy Tails and Adversarial Corruption with Near-Linear Sample Complexity

Santanu Das **Jatin Batra**
Tata Institute Of Fundamental Research, Mumbai

Abstract

Phase retrieval is the classical problem of recovering a signal $x^* \in \mathbb{R}^n$ from its noisy phaseless measurements $y_i = \langle a_i, x^* \rangle^2 + \zeta_i$ (where ζ_i denotes noise, and a_i is the sensing vector) for $i \in [m]$. The problem of phase retrieval has a rich history, with a variety of applications such as optics, crystallography, heteroscedastic regression, astrophysics, etc. A major consideration in algorithms for phase retrieval is *robustness* against measurement errors. In recent breakthroughs in algorithmic robust statistics, efficient algorithms have been developed for several parameter estimation tasks such as mean estimation, covariance estimation, robust principal component analysis (PCA), etc. in the presence of heavy-tailed noise and adversarial corruptions. In this paper, we study efficient algorithms for robust phase retrieval with heavy-tailed noise when a constant fraction of both the measurements y_i and the sensing vectors a_i may be arbitrarily adversarially corrupted. For this problem, Buna and Rebeschini (AISTATS 2025) very recently gave an *exponential* time algorithm with sample complexity $O(n \log n)$. Their algorithm needs a *robust spectral initialization*, specifically, a robust estimate of the top eigenvector of a covariance matrix, which they deemed to be beyond known efficient algorithmic techniques (similar spectral initializations are a key ingredient of a large family of phase retrieval algorithms). In this work, we make a connection between robust spectral initialization and recent algorithmic advances in robust PCA, yielding the first polynomial-time algorithms for robust phase retrieval with both heavy-tailed noise and adversarial corruptions, in fact with near-linear (in n) sample complexity.

1 INTRODUCTION

Phase retrieval. Phase retrieval is the problem of recovering a signal from *phase-less* measurements. In its simplest setting, the goal is to recover $x^* \in \mathbb{R}^n$ (up to a global sign i.e. output x minimizing $dist(x, x^*) := \min(\|x - x^*\|, \|x + x^*\|)$ from (squared) *magnitudes* of *linear* measurements

$$y_i = |\langle a_i, x^* \rangle|^2 + \zeta_i, \quad \forall i = 1, 2 \ldots, m \tag{1}$$

where y_i and a_i are known, and ζ_i denotes noise. Broadly, phase retrieval arises in inverse problems where *phase* (or sign) cannot be measured, which appears in many areas such as optics (Walther, 1963), X-ray crystallography (Harrison, 1993; Millane, 1990), microscopy (Miao et al., 2008), astronomy (Fienup and Dainty, 1987), diffraction and array imaging (Bunk et al., 2007), acoustics (Balan, 2010), blind channel estimation in wireless communications (Ranieri et al., 2013), interferometry (Demanet and Jugnon, 2017), quantum mechanics (Corbett, 2006) and quantum information (Heinosaari et al., 2013), and heteroscedastic regression (Das et al., 2023).

History of phase retrieval. While algorithms for phase retrieval in practice had been devised since the alternating minimization algorithm of Gerchberg-Saxton (Gerchberg, 1972; Fienup, 1978; Hirsch et al., 1971; Fienup, 1982; Gallagher and Liu, 1973), the problem of phase retrieval resisted provable algorithms until recently when a variety of provable techniques emerged such as spectral initialization followed by alternating minimization (Netrapalli et al., 2015; Waldspurger, 2018) or Wirtinger flow (Chen and Candes, 2015; Godeme et al., 2024; Candès et al., 2015; Zhang et al., 2017); and semi-definite programming (Candès and Li, 2014) (for more details, see Section A and also the book (Barnett et al., 2022)).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Measurement error model. In this work, we study the measurement error model where the noise ζ_i is **heavy-tailed** and homoscedastic (for the formal definition, see Model 1.2) and further an ϵ-fraction of the measurements may be **adversarially corrupted** (in particular, they can depend arbitrarily on *all* of sensing vectors a_i and responses y_i for $i \in [m]$) where $\epsilon > 0$ is known; for the formal definition, see Model 1.1 and Model 1.2. Recently, the model of strong adversarial corruption has received a lot of attention in algorithmic robust statistics, see e.g. recent breakthroughs described in the book (Diakonikolas and Kane, 2023b). The following is the notion of strong adversarial corruption in robust statistics (Diakonikolas et al., 2019; Lugosi and Neu, 2023; Diakonikolas et al., 2020):

Model 1.1. ***(Strong adversarial corruption).*** *A clean dataset $(a_i, y_i)_{i=1}^m$ is first generated according to* (1) *and then the dataset is revealed to an adversary, who can inspect all m samples and arbitrarily modify an ϵ-fraction of them. That is, the adversary is allowed to choose any subset of at most ϵm samples and replace both the covariates a_i and responses y_i in any manner. The learner only has access to the corrupted dataset provided by the adversary.*

This error model for phase retrieval was very recently studied by Buna and Rebeschini (for heavy-tailed ζ) (Buna and Rebeschini, 2025) and Dong et al. (Dong et al., 2024) (for $\zeta = 0$). Formally, for the setting of phase retrieval, in this paper, we study the following measurement models:

Model 1.2. ***(Gaussian phase retrieval with zero mean heavy tailed noise and strong adversarial corruption:)*** *Measurements y_i are generated from an unknown $x^* \in \mathbb{R}^n$ according to* (1) *with sensing vectors $a_i \sim \mathcal{N}(0, I_n)$, where ζ_i is zero-mean conditioned on a_i i.e. $\mathbb{E}[\zeta_i|a_i] = 0$ and homoscedastic i.e. its conditional fourth moment $\mathbb{E}[\zeta^4|a_i]$ and its conditional variance $\mathbb{E}[\zeta^2|a_i]$ are constants (resp. K_4^4 and σ^2) independent of a_i* [1]*. An ϵ-fraction of $\{(a_i, y_i)\}_{i=1}^m$ (where both a_i and y_i can be corrupted) are then corrupted by a strong adversary as in Model 1.1.*[2]

Model 1.3. ***(Gaussian phase retrieval with nonzero mean heavy-tailed noise and strong adversarial corruption).*** *We consider the same setting as in Model 1.2, with the modification that the noise has a nonzero conditional mean, i.e., $\mathbb{E}[\zeta_i \mid a_i] = \mu \neq 0$. Also note that μ is unknown.*

Gaussian linear measurements for phase retrieval. Gaussian measurements form the canonical and most studied setting in the phase retrieval literature (Candès and Li, 2014; Candès et al., 2015; Godeme et al., 2024; Buna and Rebeschini, 2025; Netrapalli et al., 2015; Das et al., 2023; Sun et al., 2018; Wang et al., 2018). Many influential methods, such as Truncated Wirtinger Flow, are analyzed in this model, and several applications (including heteroscedastic regression (Das et al., 2023)) naturally lead to Gaussian designs. Historically, theoretical advances in phase retrieval first emerged under Gaussian measurements and were later extended to structured models (e.g., coded diffraction patterns (CDP) (Chen and Candes, 2015)) by building on insights from the Gaussian case. In this article, we focus on the Gaussian measurement model and leave extensions to other sensing designs for future work.

Motivation for heavy-tailed noise and adversarial corruption: Real-world data are rarely clean and typically contain multiple sources of noise, broadly classified as uncorrelated noise across samples and correlated noise across the dataset. In the worst case, the former can be modelled as stochastic heavy-tailed noise, while the latter is modelled through the lens of a strong adversary (Model 1.1) that can inspect the dataset and corrupt a constant fraction of samples. Such partial corruption models are well motivated by practical applications. For instance, wireless sensor networks (Wagner, 2004; Othman et al., 2013; Tian et al., 2023; Chen and Wu, 2015) and IoT systems consist of many distributed, low-power devices that sense, collect, and transmit data, which are aggregated for monitoring, inference, and decision-making. In these large-scale deployments, typically only a small subset of sensors is compromised (Wagner, 2004; Wang et al., 2006), due to factors such as limited access, partial network penetration, or device-specific vulnerabilities. The majority of sensors remain honest; consequently, the data at the central aggregator fit a model where a fraction of measurements are adversarial and the rest are clean, resulting in a constant fraction adversarial contamination model (Model 1.1). This setting directly motivates the framework of robust statistics, which has been extensively studied; see, for example, an entire book (Diakonikolas and Kane, 2023a) is devoted to this contamination model, and numerous fundamental problems have been studied under it, including robust mean estimation (Hopkins et al., 2020), covariance estimation (Kothari et al., 2018; Li and Ye,

[1]This assumption on the second and the fourth moment are standard in robust statistics, see e.g. (Buna and Rebeschini, 2025; Oliveira and Rico, 2024; Pensia et al., 2025).

[2]As stated, the m measurements are given at once to the algorithm. However, for simplicity, we will allow our algorithms to collect a constant number of times a fresh batch of m samples as in Model 1.2, where the constant can depend on the error guarantees. This is also the assumption in prior robust iterative algorithms, e.g. see Algorithm 2 of (Buna and Rebeschini, 2025), see Step 4 of Algorithm 1 in (Liu et al., 2020), Step 4 of Algorithm 1 in (Liu et al., 2019), and the paragraph following Lemma 1 in Section 5 of (Merad and Gaïffas, 2023).

2020; Duchi et al., 2025), PCA (Kong et al., 2020; Diakonikolas et al., 2023), linear regression (Cherapanamjeri et al., 2020), and non-convex optimization (Buna and Rebeschini, 2025; Li et al., 2023). Importantly, phase retrieval itself is a highly non-convex inverse problem in which measurements are frequently acquired across multiple sensing units (Yang, 2025). In practical deployments, a fraction of sensors may be faulty or compromised (Wagner, 2004; Wang et al., 2006), leading to a corresponding fraction of corrupted measurements. While these applications provide instances of partial adversarial contamination, our goal is to emphasize that the proposed model (Model 1.2) captures both correlated corruptions and uncorrelated stochastic noise in a unified framework, abstracting more general real-world corruption.

Prior work on phase retrieval with adversarial corruptions and heavy-tailed noise: There is a long line of research on phase retrieval with adversarial corruption or heavy-tailed noise, mostly devoted to the simpler case when only the responses y_i are corrupted (and all the sensing vectors a_i are known exactly), see e.g. (Hand and Voroninski, 2016; Duchi and Ruan, 2019; Zhang et al., 2018; Qian et al., 2016; Barik et al., 2024). Our focus, in contrast, is the case where both a_i and y_i may be corrupted, which, to the best of our knowledge, is studied only very recently by Buna and Rebeschini (2025) and Dong et al. (2024). Buna and Rebeschini (2025) introduced the Model 1.2 and Model 1.3 for phase retrieval and showed exponential time algorithms that given $O(n \log n)$ samples, for a sufficiently small (but independent of n) ϵ, output x satisfying $dist(x, x^*) = O(\sigma\sqrt{\epsilon}/\|x^*\|)$[3] with high probability (for context, note that $O(\sigma\sqrt{\epsilon})$ is the information theoretically minimum error achievable in the setting of robust *mean estimation*(Diakonikolas and Kane, 2023b)). Buna and Rebeschini (2025) also describes a polynomial time algorithm for the low contamination regime where $\epsilon = O(1/n)$; however, we do not discuss that here as the low contamination regime is not the focus of our paper.

Dong et al. (2024) claimed a near-linear time algorithm with near-linear (in n) sample complexity for the noiseless ($\zeta = 0$) case with strong adversarial corruptions (Model 1.1) [4]. We discuss the techniques in both of these works in more detail in Section 2. While a complete survey of the rich algorithmic history of phase retrieval and robust statistics is beyond the scope of this work, we give a bird's eye view of other related work in Section A.

Our results: In this paper, we present the first polynomial-time algorithms (with running time $\tilde{O}(m^2 n)$) for the error models Model 1.2 and Model 1.3 that output x satisfying the error guarantee $dist(x, x^*) = O(\sigma\sqrt{\epsilon}/\|x^*\|)$ with high probability, with sample complexity $\tilde{O}(n)$ (where $\tilde{O}$ hides log factors), for any constant ϵ (*independent of* n) fraction of corruptions smaller than a constant depending only on noise to signal ratio $K_4/\|x^*\|^2$, improving the results of (Buna and Rebeschini, 2025). We also pose a question that elucidates a precise connection between robust phase retrieval in Model 1.2 and near-linear time algorithms for robust PCA, which further gives a different general perspective on the work of Dong et al. (2024) (who focused on the $\zeta = 0$ case).

We now describe our technical contribution in more detail and put it in context of prior work.

2 TECHNIQUES

Spectral initialization and gradient descent: We first briefly explain the main idea of a family of phase retrieval algorithms (Chen and Candes, 2015; Candès et al., 2015; Buna and Rebeschini, 2025; Wu and Rebeschini, 2023; Godeme et al., 2024; Ma et al., 2018; Netrapalli et al., 2015). The population risk $r(x) = \mathbb{E}[(y - \langle a, x\rangle^2)^2]/4$ for phase retrieval is nonconvex, however, it is well-known (Buna and Rebeschini, 2025; Ma et al., 2018; Chen and Candes, 2015; Candès et al., 2015; Wu and Rebeschini, 2023; Godeme et al., 2024; Netrapalli et al., 2015) that within a ball of radius $R = \Theta(\|x^*\|)$ around $\pm x^*$, $r(x)$ is smooth and strongly convex. Hence, if one could obtain an initial iterate x_0 inside this ball, *gradient descent* would converge efficiently to $\pm x^*$. It turns out that the top eigenvector of $\mathbb{E}[yaa^T]$ is parallel to x^*, and hence to obtain an initial iterate x_0, it suffices to obtain an approximate top eigenvector of $\mathbb{E}[yaa^T]$, which is typically obtained by finding the top eigenvector of an empirical estimator, such as $\frac{1}{m}\sum_i y_i a_i a_i^T$ in the simplest setup (*spectral initialization*). For several variants of spectral initialization, see (Netrapalli et al., 2015; Buna and Rebeschini, 2025; Zhang et al., 2017; Wang et al., 2018; Godeme et al., 2024; Chen and Candes, 2015).

Prior techniques for phase retrieval with strong adversarial corruption and heavy-tailed noise (Model 1.2): Naturally, the main algorithmic principle of Buna and Rebeschini (2025) and Dong et al. (2024) is to use a *robust* spectral initialization followed by a *robust* gradient descent procedure. We first de-

[3]Note that the $1/\|x^*\|$ dependence of the error is unavoidable and for a proof see Section B.

[4]While this claimed result is invalid as we discuss in more detail in paragraph titled "Comparison of our technique with prior use of truncation+filtering", it still leads to an interesting link between efficient robust PCA and robust phase retrieval; see also the discussion at the end of the second para in Section 2.

scribe their robust gradient descent procedure: they express the gradient as a mean of a random variable whose samples can be obtained from the Model 1.2, and leverage recent algorithmic breakthroughs for robust mean estimation; this idea has appeared in several recent works (Prasad et al., 2020; Li et al., 2023; Holland and Ikeda, 2019). For robust spectral initialization, Buna and Rebeschini (2025) have the insight that the top eigenvector of $\mathrm{Cov}(ya)$ also suffices for spectral initialization and use the exponential time robust covariance estimator (Oliveira and Rico, 2024) to estimate $\mathrm{Cov}(ya)$, and find the top eigenvector of the estimate. On the other hand Dong et al. (2024) (who focuses on the $\xi = 0$ case) showed that given sufficient samples, robustly finding the top eigenvector of $\mathbb{E}[yaa^T]$ can be reduced to a convex Ky-fan-2 norm optimization problem, which can be converted into a packing semidefinite program for which nearly-linear time solvers exist. [5]

2.1 Our technical contribution:

Robust phase retrieval for zero mean noise. While we still use the same robust gradient descent procedure as prior works (Buna and Rebeschini, 2025; Dong et al., 2024), we show that the result of Kong et al. (2020) for robust PCA can be leveraged to obtain, in polynomial time, a robust spectral initialization in the Model 1.2 that suffices for initializing the robust gradient descent procedure. Kong et al. (2020) shows that for *bounded* random variables X satisfying a certain fourth-moment condition, the top eigenvector of the covariance matrix $\mathrm{Cov}(X)$ can be approximated in polynomial time with sample complexity $\tilde{O}(n)$. We show that a well-chosen *truncation* on ya leads to an appropriately bounded random variable. While truncation combined with filtering is classical, it is not a priori clear that a valid truncation exists for robust phase retrieval: over-truncation distorts the distribution, and under-truncation weakens filtering. It turns out that with the bounded fourth-moment assumption on the noise, our choice of truncation gives a sufficiently good approximation to the top eigenvector of $\mathrm{Cov}(ya)$ in polynomial time with sample complexity $\tilde{O}(n)$ by employing the result of Kong et al. (2020). Hence, together with a robust gradient descent procedure ((Buna and Rebeschini, 2025) or (Dong et al., 2024)), this leads to a polynomial time algorithm for the error Model 1.2. Specifically, we show the following theorem (informally stated, whose formal version appears in Theorem 4.1):

[5] Their reduction for near-linear sample complexity appears to have a bug; however, their reduction works if the sample complexity is allowed to be more than $\tilde{O}(n)$ (but still a polynomial in n), we elaborate on this in Section 4.2.

Theorem 2.1. ***(Informal.)*** *Consider the Model 1.2. Assume an upper bound r_{up} on $K_4/\|x^*\|^2$ is known. There exists a polynomial time algorithm (Algorithm 1 (this work) +Algorithm 2 from (Buna and Rebeschini, 2025)) such that if the corruption fraction $\epsilon \leq C_1/(r_{up}^2+1)^2$ and the number of samples $m \geq C_2 n(r_{up}^2+1)\log(2n/(\delta\epsilon))/\epsilon$ (where C_1 and C_2 are absolute constants), then with probability at least $1-\delta$, the output x_{out} of the algorithm satisfies*

$$dist(x_{out}, x^*) \leq O\left(\frac{\sigma}{\|x^*\|}\sqrt{\epsilon}\right).$$

Comparison of our work with Buna and Rebeschini (2025). Since Buna and Rebeschini (2025) studies the same problem as ours, we compare the results here. Convergence guarantees for robust iterative optimization algorithms typically rely on strong convexity and smoothness (Prasad et al., 2020), which hold only locally in non-convex problems. Hence, robust non-convex methods often require a good initialization to ensure that the iterates enter a well-behaved region (methods with random initializations often don't have as good sample complexity guarantees in the robust setting (Li et al., 2023)). Phase retrieval (PR) is a canonical example of such a problem. Buna and Rebeschini (2025) (AISTATS 2025) showed that the traditional spectral-initialization method for PR to land in a strongly convex and smooth region around the minimum, can be interpreted as computing the top eigenvector of $\mathrm{Cov}(ya)$, enabling the use of robust PCA methods for the initialization. Starting from such an initialization, Buna and Rebeschini (2025) analyzed the robust gradient-descent procedure for phase retrieval and showed that the hyperparameters (such as the learning rate, sample complexity, and allowable corruption level) can be chosen so that the iterates of the non-convex PR objective remain within the strongly convex and smooth region established by the initializer. Robust gradient descent is well understood given strong convexity and smoothness at the iterate, and methods using second order information for non-convex landscapes are also known e.g. the definition of robust gradient in Buna and Rebeschini (2025) appears for the first time in Prasad et al. (2020), see also Cherapanamjeri et al. (2020) utilizing this idea, see Li et al. (2023) for robust non-convex optimization using second-order information at the cost of higher sample complexity. The key and poorly understood component in sample-efficient robust non-convex procedures is therefore robust initialization, Buna and Rebeschini (2025) gave the first such (spectral) initialization for a non-convex problem, which, however, needed exponential time. Our contribution is to show that one can find a choice of truncation on the covariates that, together with the robust PCA algorithm (filtering) of

Kong et al. (2020) gives a polynomial-time spectral initializer with near-linear sample complexity that tolerates a constant fraction of corruptions, thereby improving the result of Buna and Rebeschini (2025) from exponential time to polynomial time. The technical difficulty of finding such truncation + filtering combinations is explained below.

Comparison of our technique with prior use of truncation+filtering. While truncation and filtering are classical tools in robust statistics, their successful interplay is highly problem-dependent. Several examples illustrate this point: for instance, Diakonikolas et al. (2022)'s robust sparse mean estimation proceeds by projecting (i.e. truncating) means to an ℓ_∞ ball around the median-of-means estimator, followed by the filtering procedure of Balakrishnan et al. (2017). Catoni and Olivier Catoni (2012) use a customized truncation-based estimator for heavy-tailed mean estimation. The work Sasai (2022) addresses the robust sparse regression problem by first using coordinate-wise truncation together with a filtering step to remove adversarially corrupted design vectors, and then they perform regression by minimizing the Huber loss, which is essentially a truncated ℓ_2 loss, to handle corruption in the responses. On the other hand, the proof attempt in the (incorrect) submission Dong et al. (2024) uses truncation together with concentration inequalities, which is conceptually similar to the truncation-plus-filtering framework, where the validity of their argument appears to rely critically on the last inequality on page 14 of their manuscript, which claims that $\Pr_{x\sim\mathcal{N}(0,1)}(x^4 \geq \ln z/c) \leq z^{-3}$ holds for all $z > 1$ for some constant c; however, this is false, and more generally the integral $\int_{\tilde{\tau}}^{\infty} \Pr_{x\sim\mathcal{N}(0,1)}(x^4 \geq \ln z/c)dz$ diverges for all $\tilde{\tau}$ and constants $c > 0$; from this viewpoint, the issues in their argument can be interpreted as a failed attempt to apply a truncation-based method, whereas in a concurrent work Dong et al. (2025), the authors of Dong et al. (2024) claim a result that resolves the incorrect part of their earlier results for phase retrieval in the zero-noise setting using a sophisticated truncation+filtering techniques. These examples highlight that the correct use of truncation and filtering is subtle and depends critically on the specifics of the problem: over-truncation can destroy essential structure, while under-truncation can render filtering ineffective. We truncate the vectors ya using a norm-based cutoff determined by the input parameters, chosen large enough to keep the top eigenvector of the covariance of the truncated ay close to the top eigenvector of the true covariance, yet small enough to apply the robust PCA method of Kong et al. (2020). Furthermore, unlike earlier truncation-based phase retrieval methods—which typically truncate only on the responses—our method truncates both the responses and the sensing vectors. We also show that our method remains valid even when the noise has a nonzero mean, which requires robustly solving a special case of blind deconvolution. Our work provides a principled and effective initialization scheme using truncation + filtering that may naturally extend to other robust nonconvex estimation settings.

Towards near-linear time algorithms for robust phase retrieval: In fact, we can show that the appropriately truncated empirical distribution of ya over m samples satisfies a certain stability condition (Definition 4.6, (Diakonikolas et al., 2023)) that can be used to employ the nearly-linear time PCA algorithm of Diakonikolas et al. (2023) for robust spectral initialization, at the cost of losing the $\widetilde{O}(n)$ sample complexity; however, the sample complexity m still remains polynomial in n (even for heavy-tailed noise). This also provides a different perspective on Dong et al. (2024), whose reduction to convex Ky-fan-2 norm optimization relies on the same stability condition (see Definition 4.6), for the zero-noise case. We pose the open problem: does the empirical distribution over $\widetilde{O}(n)$ appropriately truncated samples of ya satisfy stability (Definition 4.6)? This question, if answered affirmatively, would imply a near-linear time algorithm for robust phase retrieval in the Model 1.2.

Robust phase retrieval for non-zero mean noise: Next, we study Model 1.3, which considers the case of non-zero unknown mean noise. (Buna and Rebeschini, 2025) introduced a *symmetrization-based* trick, also used in (Pensia et al., 2025), that converts phase retrieval with non-zero mean noise into a special case of the robust blind deconvolution problem. Similar to Model 1.2, Buna and Rebeschini (2025) provided an exponential-time algorithm for Model 1.3 with the same guarantees as those given for Model 1.2. We show that our techniques can also be extended to Model 1.3. However, estimating $\|x^*\|$ in this setting is not straightforward. Refer Remark G.5 for more details, in fact, the argument in Buna and Rebeschini (2025) contains a gap in the robust estimation of $\|x^*\|$. In this work, we provide a fix for this issue by first obtaining a crude robust estimate of $\|x^*\|$, using it to carefully choose the truncation parameter, and then using robust PCA to fine-tune the previous crude estimate. Specifically, we show the following theorem (informally stated, whose formal version appears in Theorem 5.1):

Theorem 2.2. ***(Informal.)*** *Under Model 1.3 and the same settings as in Theorem 2.1, there exists a polynomial-time algorithm (Algorithm 2 (this work) + Algorithm 4 from (Buna and Rebeschini, 2025)) that achieves the same guarantees as in Theorem 2.1, except that the requirement on m changes to $m \geq$*

$C_2\, n\, (r_{up}^2+1)^2\, \frac{\log\left(2n/(\delta\epsilon)\right)}{\epsilon}$.

Roadmap of the paper. In Section 3, we give the necessary preliminaries for algorithmic robust statistics. In Sections 4 and 5, we describe our polynomial-time algorithms with $\tilde{O}(n)$ sample complexity for both models 1.2 and 1.3 respectively, and prove that the output x satisfies $dist(x, x^*) = O(\sigma\sqrt{\epsilon}/\|x^*\|)$. The crux of the argument for both models 1.2 and 1.3, presented in Sections 4.1 and G.3, is that an appropriate truncation ensures the robust PCA estimator of (Kong et al., 2020) suffices for spectral initialization. In Section 4.2, we discuss the connection to (Dong et al., 2024) and pose an open question whose affirmative answer yields a nearly-linear time algorithm for robust phase retrieval in Model 1.2. Finally, Section 6 concludes with open problems.

3 PRELIMINARIES

Notations used in paper. We use $\|x\|$ for the ℓ_2 norm of a vector x, $\|A\|_{\rm op}$ for the operator norm of a matrix A, and $\|A\|_F$ for its Frobenius norm. We write $\mathcal{B}(x, R)$ for the ℓ_2 ball in $\mathbb{R}^n$ centered at x with radius R. As is standard in the statistics literature, we slightly abuse notations by using the same symbols for a random variable and its realization when clear from context. We denote unspecified constants by $C_1, C_2, \ldots$, where C_i may vary across results.

We now state the main results in algorithmic robust statistics that our paper relies on.

3.1 Robust mean estimation

For robust mean estimation, we will employ the following results of Diakonikolas et al. (2020) and Hopkins et al. (2020).

Theorem 3.1. *[Theorem 1.4, Proposition 1.5, Proposition 1.6 in ((Diakonikolas et al., 2020))] Suppose a dataset S of size m is sampled i.i.d. from a distribution $\mathcal{D}$ with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. Let T denote an $\epsilon-$corrupted version of S, where the corruption is performed by a strong adversary (i.e. up to ϵm points in S are arbitrarily modified to obtain T). For ϵ less than a sufficiently small universal constant, there exists a polynomial time algorithm on input T, ϵ that efficiently computes $\widehat{\mu}$ such that, with probability at least $1-\delta$,*

$$\|\widehat{\mu}-\mu\| = \sqrt{\|\Sigma\|_{\rm op}}\, O\left(\sqrt{\frac{{\rm r_{eff}}(\Sigma)}{m}} + \sqrt{\epsilon} + \sqrt{\frac{\log(1/\delta)}{m}}\right), \tag{2}$$

where ${\rm r_{eff}}(\Sigma) = {\rm tr}(\Sigma)/\|\Sigma\|_{\rm op}$ is the effective rank of matrix Σ. The polynomial time algorithm is completely oblivious to Σ (Theorem A.3 in (Diakonikolas et al., 2020)). Let's denote the time complexity of this polynomial-time algorithm by $T_{\text{rob-mean}}(m, n)$. Further, by (Hopkins et al., 2020), if an upper bound σ_{up} on $\|\Sigma\|_{op}$ is known, the above guarantee in (2) can be obtained in time $\tilde{O}(mn)$ (with $\|\Sigma\|_{op}$ replaced with σ_{up} in (2)).

3.2 Robust principal component analysis

For robust principal component analysis (PCA), we will employ the following results of (Kong et al., 2020).

Theorem 3.2. *(Proposition 2.6 from (Kong et al., 2020).) Let $S = \{X_i\}_{i=1}^m$ be drawn i.i.d from a distribution $\mathcal{D}$ supported on $\mathbb{R}^n$. Let's define the* second-order raw moment *of $X \sim \mathcal{D}$ as $\Sigma_r := \mathbb{E}_{X\sim\mathcal{D}}[XX^\top]$. Given $\delta \in (0, 0.5)$, and a corrupted dataset T such that an $\epsilon \in (0, 1/36]$ fraction of the points are corrupted arbitrarily, suppose:*

- *The distribution $\mathcal{D}$ has bounded support such that $\|XX^\top - \Sigma_r\| \le B$ for all $X \sim \mathcal{D}$ with probability one.*
- *The distribution $\mathcal{D}$ has bounded fourth moments, i.e.,*

$$\max_{x\in\mathbb{R}^n:\|x\|\le 1} \mathbb{E}_{X\sim\mathcal{D}}\left[\langle xx^\top, XX^\top - \Sigma_r\rangle^2\right] \le v'^2.$$

Then, with probability at least $1-\delta$, there exists a polynomial-time algorithm (Algorithm 2 from (Kong et al., 2020)) that efficiently computes $\hat{u} \in \mathbb{R}^{n\times 1}$ and $\lambda_{\hat{u}} \in \mathbb{R}$ such that it satisfies:

$$\mathrm{Tr}[P_1(\Sigma_r)] - \mathrm{Tr}[\hat{u}^\top(\Sigma_r)\hat{u}] = O\left(\epsilon\cdot \mathrm{Tr}[P_1(\Sigma_r)] + \sqrt{\epsilon}\cdot v'\right), \tag{3}$$

$$|\lambda_{\hat{u}} - \mathrm{Tr}[P_1(\Sigma_r)]| \le O\left(\epsilon\cdot \mathrm{Tr}[P_1(\Sigma_r)] + \sqrt{\epsilon}\cdot v'\right) \tag{4}$$

provided that

$$m = \Omega\left((n + (B/v')\sqrt{\epsilon})\log(n/(\delta\epsilon))/\epsilon\right).$$

Here, $P_k(\cdot)$ is the best rank-k approximation of a matrix in ℓ_2. The time complexity of Algorithm 2 from (Kong et al., 2020) is $\widetilde{O}(m^2 n)$.

Remark 3.3. Algorithm 2 from Kong et al. (2020) is a robust rank-one PCA routine. Given $T = \{X_j X_j^\top\}$, it applies a filtering step (Algorithm 3 in Kong et al. (2020), "Double Filtering") to remove adversarial elements, averages the remaining matrices to form $\hat{M}$, and returns the top eigenvector $\hat{u}$ of $\hat{M}$, satisfying Eq. (3). Our implementation also computes the corresponding leading eigenvalue (an estimate of the largest eigenvalue of Σ_r). Although this step is not explicitly performed in Kong et al. (2020), its guarantee (Eq. (4)) follows directly from the analysis in that work. See Section D.1 for the proof of Equation (4).

4 ROBUST PHASE RETRIEVAL WITH ZERO MEAN NOISE

In this section, we first present the formal version of Theorem 2.1, and then progressively build up the necessary tools and results required to prove it.

Theorem 4.1. *Assume that an upper bound r_{up} on $K_4/\|x^*\|^2$ is known. There exists a polynomial-time algorithm (namely, Algorithm 1 combined with Algorithm 2 from (Buna and Rebeschini, 2025)) such that, if the total number of samples is $m = m_1 + m_2 + P\tilde{m}$, and the following conditions hold: $m_1 \geq C_2 \log(2/\delta)(1 + r_{up}^2), m_2 \geq C_3 n(r_{up}^2 + 1)\frac{\log(2n/(\delta\epsilon))}{\epsilon}$, $\epsilon \leq \frac{C_1}{(r_{up}^2+1)^2}$, $\tilde{m} \geq C_4 \cdot \max\{n \log n, \log(1/\delta)\} \cdot r_{up}^2$, then, with probability at least $1 - (P + 1)\delta$, the algorithm outputs x_P satisfying*

$$\frac{dist(x_P, x^*)}{\|x^*\|} \leq C_5 \exp\left(-C_6 P(1 - \sqrt{\epsilon})\right) + C_5 \frac{\sigma}{\|x^*\|^2}\left(\sqrt{\frac{n \log n}{\tilde{m}}} + \sqrt{\frac{\log(1/\delta)}{\tilde{m}}} + \sqrt{\epsilon}\right) \quad (5)$$

Moreover, the time complexity of the algorithm is $O(T_{\text{rob-mean}}(m_1, 1) + m_2^2 n + PT_{\text{rob-mean}}(\tilde{m}, n))$.

Using the fact that ${}^{\sigma}/_{\|x^*\|^2} \leq {}^{K_4}/_{\|x^*\|^2} \leq r_{up}$, we conclude that for $P = \Omega\left(({}^{1}/_{1-\sqrt{\epsilon}}) \log\left({}^{1}/_{r_{up}\sqrt{\epsilon}}\right)\right)$, we have $\text{dist}(x_P, x^*) = O\left({}^{\sigma\sqrt{\epsilon}}/_{\|x^*\|}\right)$. The sample complexity and corruption level for both robust spectral initialization and gradient descent remain of the same order for $\sigma = 0$ and $\sigma \neq 0$, as follows by combining our analysis with that of (Buna and Rebeschini, 2025). We will use the following result of (Buna and Rebeschini, 2025) who showed that there exists a polynomial time algorithm (robust gradient descent) with sample complexity $\tilde{O}(n)$ for robust phase retrieval if given a point x_0 inside a ball of radius $\|x^*\|/9$ centered at $\pm x^*$.

Theorem 4.2. ***(Gradient descent for phase retrieval, see Theorem 3.3 from Buna and Rebeschini (2025).)*** *There exists a polynomial time algorithm (see Algorithm 2 from (Buna and Rebeschini, 2025)) that, given a point $x_0 \in B(\pm x^*, \|x^*\|/9)$ and if $\epsilon \leq \frac{C_1\|x^*\|^4}{\sigma^2}$ and P draws of $\tilde{m}$ samples where $P\tilde{m} \geq C_2 P \cdot \max\{n \log n, \log(1/\delta)\} \cdot \frac{\sigma^2}{\|x^*\|^4}$, with probability at least $1 - P\delta$, outputs x_P satisfying*

$$\frac{dist(x_P, x^*)}{\|x^*\|} \leq C_3 \exp\left(-C_4 P(1 - \sqrt{\epsilon})\right) + C_3 \frac{\sigma}{\|x^*\|^2}\left(\sqrt{\frac{n \log n}{\tilde{m}}} + \sqrt{\frac{\log(1/\delta)}{\tilde{m}}}\right) + C_3 \frac{\sigma}{\|x^*\|^2}\sqrt{\epsilon}.$$

Remark 4.3. Note that (Buna and Rebeschini, 2025) and prior works (Liu et al., 2020, 2019; Merad and Gaïffas, 2023) use fresh samples at each step of robust gradient descent (see Footnote 2 on page 2). This is standard because drawing multiple batches of fresh samples is essentially similar to random sample splitting. Specifically, we can take $M = O(m'(P+2))$ samples from Model 1.2 at the beginning and randomly divide them into $P + 2$ batches. By Claim C.1, if $m' = \Omega((\log(P + 2) + \log(1/\delta))/\varepsilon^2)$, then with probability at least $1 - \delta$, every batch has at most a 2ε fraction of corrupted samples. For a formal treatment, see Section C. We can use 2 batches for spectral initialization and the remaining P batches for gradient descent. Under sample splitting, all theorem statements and proofs in our work and Buna and Rebeschini (2025) remain the same, except ε is replaced by 2ε and the sample complexity is replaced by maximum of sample complexity appears in our work (and (Buna and Rebeschini, 2025)) and $c_{\varepsilon,r_{up}} = 2\max\{\log(1/\delta), \log\left({}^{1}/_{1-\sqrt{\epsilon}} \log\left({}^{1}/_{r_{up}\sqrt{\epsilon}}\right)\right)\}/\varepsilon^2$, where $c_{\varepsilon,r_{up}}$ is a constant depending on problem parameters and ϵ. For simplicity, we do not take this path. We leave it as an open problem whether robust gradient descent can be performed using only a single batch of samples.

In the following Section 4.1, we show how to obtain such a point x_0 which gives the desired algorithm together with Theorem 4.2.

4.1 Robust Spectral initialization

Define the random variable $X := ya$ and denote its covariance matrix as $\Sigma := \mathbb{E}[XX^T]$ (note that $\mathbb{E}[X] = 0$). We begin with the idea of Buna and Rebeschini (2025), which interprets the spectral initialization of a long line of work (Netrapalli et al., 2015; Chen and Candes, 2015) as follows: the top eigenvector of the covariance matrix $\Sigma = \mathbb{E}[XX^T]$ (note that $\mathbb{E}[X] = 0$) is parallel to x^*:

$$\mathbb{E}[y^2 aa^T] = \mathbb{E}\left(\langle a, x^*\rangle^4 aa^T\right) + \mathbb{E}\left(aa^T\, \mathbb{E}[\zeta^2 \mid a]\right) \overset{(a)}{=} (3\|x^*\|^4 + \sigma^2)\mathbb{I}_n + 12\|x^*\|^2 x^* x^{*T},$$

with eigenvalue $15\|x^*\|^4 + \sigma^2$ (where (a) follows from the Lemma H.1).

This implies that it is sufficient to estimate only the top eigenvector of the matrix Σ, as we can then scale this direction to obtain an initial estimate of x^*. Importantly, there is no need to estimate the full covariance matrix Σ.

Our tool for this purpose is the robust PCA algorithm of Theorem 3.2 which estimates the top eigenvector with sample complexity $\tilde{O}(n)$ for *bounded* random variables satisfying a certain fourth-moment condition. $X = ya$ is unbounded; however, one can construct a bounded random variable by *truncating* X at a norm cutoff τ to be chosen later, without affecting the covariance matrix by much (due to Chebyshev-like concentration because X also has a bounded fourth moment). Formally, we show:

Lemma 4.4. ***(Truncation.)*** *Define $\widetilde{X} := X\mathbb{I}_{\|X\|\leq\tau}$, $\widetilde{\Sigma}_r := \mathbb{E}[\widetilde{X}\widetilde{X}^\top]$, where τ is the truncation parameter. Then, the following holds:*

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\text{op}} = O(n(K_4^2 + \|x^*\|^4)^2/\tau^2). \tag{6}$$

The proof of the Lemma 4.4 can be found in Section E.1.

Estimating the error due to truncation. It turns out that for estimating the top eigenvector of Σ, we can afford a truncation of $\tau = O(\sqrt{n}((K_4^2/\|x^*\|^4) + 1)(\|x^*\|^2))$ because the top eigenvector of $\widetilde{\Sigma}_r$ only needs to lie in a sufficiently small ball around x^*. Specifically, the above choice of τ gives $\|\Sigma - \widetilde{\Sigma}_r\|_{\text{op}} \leq \frac{\sqrt{2}\|x^*\|^4}{9}$. Let u be the normalized top eigenvector of $\widetilde{\Sigma}_r$, then together with Theorem H.3 (observing that $\lambda_1(\Sigma) - \lambda_2(\Sigma) = 12\|x^*\|^4$), we get that

$$dist(u, x^*/\|x^*\|) \leq \frac{2\sqrt{2}\|\Sigma - \widetilde{\Sigma}_r\|_{\text{op}}}{12\|x^*\|^4} \leq \frac{1}{27}.$$

If the norm $\|x^*\|$ was known and u could be estimated exactly, then the above calculation shows that the estimate $\|x^*\| \cdot u$ satisfies $dist(\|x^*\| \cdot u, x^*) \leq \|x^*\|/27$, which suffices as the input x_0 to the algorithm of Theorem 4.2. This suggests that we first estimate $\|x^*\|$ and use this estimate to define the truncation parameter. Use this parameter to truncate the r.v. ya, and then apply Theorem 3.2 to obtain the direction for x^*, and finally, combine the direction and scaling to get the final estimate of x^*. See Algorithm 1 for the full algorithm. The explicit constant in the truncation parameter (1100) follows from constants appearing in Kong et al. (2020) and Lemma H.2. We now present the main theorem analyzing Algorithm 1 whose proof can be found in Section E.2.

Theorem 4.5. ***(Robust spectral initialization for phase retrieval.)*** *Under the setting and notation of Algorithm 1, if $\epsilon \leq C_1(r_{up}^2+1)^{-2}$, $m_1 \geq \log(2/\delta)(1 + r_{up}^2)$ and $m_2 \geq C_3\left(n(r_{up}^2+1)\log(2n/(\delta\epsilon))/\epsilon\right)$. Then, with probability at least $1-\delta$, $dist(x_0, x^*) \leq \|x^*\|/9$. The time complexity of Algorithm 1 is $O(T_{\text{rob-mean}}(m_1, 1) + m_2^2 n)$.*

Theorem 4.1 now follows by combining Theorem 4.2 and Theorem 4.5.

4.2 Connection between stability-based robust PCA and robust phase retrieval

In this section, we show a connection (Proposition 4.7) between near-linear time stability-based robust PCA algorithms and robust spectral initialization. Diakonikolas et al. (2023) studied the robust PCA problem and gave a nearly linear time robust PCA algorithm (Algorithm 2 (SampleTopEigenvector)) given that the empirical uniform distribution over good samples satisfies the stability condition. The stability condition from Diakonikolas et al. (2023) is as follows:

Algorithm 1 Spectral Initialization.

Inputs: failure probability $\delta \in (0,1)$, corruption fraction $\epsilon > 0$, an upper bound r_{up} on $K_4/\|x^*\|^2$, access to 2 batches of samples of sizes $m_1 = \Omega(\log(2/\delta)(1+r_{up}^2))$ and $m_2 = \Omega\left(n(r_{up}^2+1)\log(2n/(\delta\epsilon))/\epsilon\right)$ respectively from Model 1.2.

Output: $x_0 \in \mathbb{R}^n$

1: $\|x^*\|$ **calculation**:
- Receive a set of samples $T_y = \{(y_j)\}_{j=1}^{m_1}$ of size m_1 from the robust phase retrieval Model 1.2.
- Let $\hat{\|x^*\|}^2$ be the maximum of zero and the robust mean estimate of the dataset T_y by the Algorithm in Theorem 3.1.
- Calculate $\hat{\|x^*\|} = \sqrt{\hat{\|x^*\|}^2}$ as robust estimate of $\|x^*\|$.

2: **Robust PCA**:
- Calculate the truncation parameter $\hat{\tau} = 1100\sqrt{n}(r_{up}^2+1)\hat{\|x^*\|}^2$.
- Receive a set of samples $T = \{(a_j, y_j)\}_{j=1}^{m_2}$ of size m_2 from the robust phase retrieval Model 1.2.
- Form the truncated dataset $\widetilde{D} = \{\widetilde{X}_j = X_j\mathbb{I}_{\|X_j\|\leq\hat{\tau}} : X_j = y_j a_j\}_{j=1}^{m_2}$.
- Let $\hat{u}$ be the output of the robust PCA algorithm (Algorithm 2 from (Kong et al., 2020)) applied to the truncated dataset $\widetilde{D}$.

3: **Scaling**: Return $x_0 = \hat{\|x^*\|}\hat{u}$.

Definition 4.6. *(See Definition 2.12 from (Diakonikolas et al., 2023).)* ***Stability Condition.*** *Let $0 < \epsilon < 1/2$ and $\epsilon \leq \gamma < 1$. A distribution G on $\mathbb{R}^d$ is called (ϵ, γ)-stable with respect to a PSD matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, if for every weight function $w : \mathbb{R}^d \to [0,1]$ with $\mathbf{E}_{X\sim G}[w(X)] \geq 1-\epsilon$, the weighted second moment matrix, $\mathbf{\Sigma}_{G_w} := \mathbf{E}_{X\sim G}\left[w(X)XX^\top\right]/\mathbf{E}_{X\sim G}[w(X)]$, satisfies that $(1-\gamma)\mathbf{\Sigma} \preceq \mathbf{\Sigma}_{G_w} \preceq (1+\gamma)\mathbf{\Sigma}$.*

The Algorithm 2 (SampleTopEigenvector) from Diakonikolas et al. (2023) (whose detailed gurantees can be found in Section F.1) returns u with $u^\top\mathbf{\Sigma}u \geq (1-O(\gamma))\|\mathbf{\Sigma}\|_{\text{op}}$ in time $O\left(\frac{m'n}{\gamma^2}\log^4(n/\epsilon)\right)$, where m' is the no of input samples given that the uncoruupted samples satisfies Definition 4.6.

Proposition 4.7. ***(Informal.)*** *If the truncated empirical distribution of size m satisfies the stability con-*

dition (Definition 4.6), then we obtain an algorithm for robust phase retrieval with runtime $\widetilde{O}(mn)$, sample complexity $\widetilde{O}(m)$, and contamination tolerance $\epsilon \leq C_2/(1+r_{up}^2)$. The algorithm is A_{alt}+ Algorithm 2 from (Buna and Rebeschini, 2025), where A_{alt} is identical to Algorithm 1 except that Step 2 replaces Algorithm 2 of (Kong et al., 2020) with SampleTopEigenvector (Algorithm 2) from Diakonikolas et al. (2023).

See Section F.2 for the formal statement and proof of Proposition 4.7. Such a proposition is also implicit in the work of Dong et al. (2024). Dong et al. (2024) claimed a *near-linear* time for the phase retrieval problem in the zero-noise setting with adversarial corruption (which is invalid, as discussed in Footnote 4). They reduced the problem of robust spectral initialization (their Lemma 4.1) to an optimization problem involving the Ky Fan 2 norm problem, which can be solved in *linear time* by packing SDP solvers. However, their reduction crucially relies on their stability-based lemma (Lemma 4.2).

By Proposition 4.7, we conclude that to obtain nearly-linear time algorithms with sample complexity $\widetilde{O}(n)$ for robust phase retrieval, it suffices to prove the stability condition (Definition 4.6) for the empirical distribution with $\widetilde{O}(n)$ truncated samples, specifically, to answer in the affirmative the following question:

Question 4.8. Let τ be $\Theta(\sqrt{n})$. Can we prove or disprove that the truncated empirical distribution of $Y = (\langle a, x^*\rangle^2 + \zeta)a\mathbb{I}_{\|(\langle a,x^*\rangle^2+\zeta)a\|\leq\tau}$ over $\widetilde{O}(n)$ samples satisfies Definition 4.6 with respect to $\Sigma := \|x^*\|^2\mathbb{I}_n + 2x^*x^{*\top}$?

To answer the question partly, we show that the empirical distribution over $\widetilde{O}(n^5)$ samples of the random variable Y satisfies the stability definition Definition 4.6. We prove this formally in Lemma F.3. Lemma F.3 leads to the Theorem F.5 (which can be found with proof in Section F.4), which shows that the algorithm A_{alt} suffices for robust spectral initialization (which, along with Theorem 4.2 gives an algorithm for robust phase retrieval) and runs in $\widetilde{O}(T_{\text{rob-mean}}(m_1, 1) + m_2 n)$ time but at the cost of higher sample complexity $m_0 = m_1 + m_2 = \Omega(n^5)$.

5 ROBUST PHASE RETRIEVAL WITH NON-ZERO MEAN NOISE

In this section, we study Model 1.3. Using the symmetrization trick from Buna and Rebeschini (2025), we reduce phase retrieval to a special case of blind deconvolution (Buna and Rebeschini, 2025). We then present the formal version of Theorem 2.2. For a discussion on the statistical challenges (Remark G.5) and the tools we develop to address them, see paragraph titled "Our approach" and two paragraphs below the proof of Lemma G.6 in Section G.3.

When the noise has an *unknown* (possibly non-zero) mean, the approach in Algorithm 1 becomes invalid, since $\mathbb{E}[y] = \|x^*\|^2 + \mu$, where μ is unknown. To address this issue, we follow the idea of *symmetrization-based pre-processing* from Buna and Rebeschini (2025). This idea has been used before, by Pensia et al. (2025) for robust linear regression. Assume we are given two independent observations: $y = (\langle a^*, x^*\rangle)^2 + z, \quad y' = (\langle a^{*'}, x^*\rangle)^2 + z'$, where z, z' are independent noise terms with unknown mean. Define: $\upsilon := {y-y'}/2, \quad b := {a^*+a^{*'}}/\sqrt{2}, \quad c := {a^*-a^{*'}}/\sqrt{2}, \quad \zeta := \frac{z-z'}{2}$. Using the identity $(\langle a^*, x^*\rangle)^2 - (\langle a^{*'}, x^*\rangle)^2 = \langle a^* + a^{*'}, x^*\rangle \cdot \langle a^* - a^{*'}, x^*\rangle$, we can rewrite the above as and obtain a new model of the form: $\upsilon = \langle b, x^*\rangle \cdot \langle c, x^*\rangle + \zeta$. where: $b, c \sim \mathcal{N}(0, I_d)$ are independent, $\mathbb{E}[\zeta \mid b, c] = 0, \mathbb{E}[\zeta^2 \mid b, c] = \sigma^2/2, \mathbb{E}[\zeta^4 \mid b, c] \leq 2K_4^4$. This is called the blind deconvolution problem. Note that the population loss $r(x) := \mathbb{E}[(\upsilon - \langle b, x\rangle\langle c, x\rangle)^2]/2$ is strongly convex and smooth inside a small ball around x^* (see Proposition G.2 and Proposition G.3), implying the success of robust spectral initialization + robust gradient descent. We now state the formal version of Theorem 2.2, whose proof is in Section G.5.

Theorem 5.1. *Consider the same settings as Theorem 4.1. There exists a polynomial-time algorithm (namely, Algorithm 2 (ours work) combined with Algorithm 4 (robust gradient descent) from (Buna and Rebeschini, 2025)) such that, if the total number of samples is $m = 2m_1 + 2m_2 + 2P\tilde{m}$, and the following conditions hold: $m_1 \geq C_2 \log(2/\delta)(1 + r_{up}^2), m_2 \geq C_3 {n(r_{up}^2+1)^2 \log(2n/(\delta\epsilon))}/{\epsilon}, 2\epsilon \leq {C_1}/{(r_{up}^2+1)^2}, \quad \tilde{m} \geq C_4 \cdot \max\{n\log n,\ \log(1/\delta)\} \cdot r_{up}^2$, then, with probability at least $1 - (P + 1)\delta$, the algorithm outputs x_P satisfying Equation* (5). *Moreover, the time complexity of the algorithm is $O(T_{\text{rob-mean}}(m_1, 1) + m_2^2 n + PT_{\text{rob-mean}}(\tilde{m}, n))$.*

6 CONCLUSION

We establish a connection between robust phase retrieval and robust PCA, yielding the first polynomial-time algorithms that tolerate a constant corruption fraction depending on the noise-to-signal ratio (independent of dimension), and extend this approach to non-zero mean noise via robust blind deconvolution. An open question is whether the corruption level can be made independent of the noise-to-signal ratio; the known information-theoretic lower bound is $O(\sigma\epsilon)$ for robust Gaussian linear regression (Bakshi and Prasad, 2021), which extends to phase retrieval for $\|x^*\| = 1$, and whether a sharper bound exists or an algorithm achieving error $O(\sigma\epsilon/\|x^*\|)$ remains open.

7 Acknowledgment

We thank Piyush Srivastava and the AISTATS reviewers for insightful discussions that helped strengthen our work and place it in context. We thank Pravesh Kothari for useful discussions during the authors' visit to the International Centre for Theoretical Sciences, Tata Institute of Fundamental Research (ICTS-TIFR), Bangalore, for the conference GEOMETRY, PROBABILITY, AND ALGORITHMS. We also acknowledge the support of the Department of Atomic Energy, Government of India, under Project No. RTI4014.

References

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212. PMLR, 2017.

Radu Balan. On signal reconstruction from its spectrogram. In *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–4, 2010. doi: 10.1109/CISS.2010.5464828. URL `https://ieeexplore.ieee.org/document/5464828?signout=success`.

Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2005.07.001. URL `https://www.sciencedirect.com/science/article/pii/S1063520305000667`.

Adarsh Barik, Anand Krishna, and Vincent YF Tan. A sample efficient alternating minimization-based algorithm for robust phase retrieval. *arXiv:2409.04733*, 2024. URL `https://arxiv.org/abs/2409.04733`.

Alexander H. Barnett, Charles L. Epstein, Leslie Greengard, and Jeremy Magland. *Geometry of the Phase Retrieval Problem Graveyard of Algorithms*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2022. ISBN 9781009008556.

Alex Buna and Patrick Rebeschini. Robust gradient descent for phase retrieval. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2080–2088. PMLR, 03–05 May 2025. URL `https://proceedings.mlr.press/v258/buna25a.html`.

Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K. Satapathy, and J. Friso van der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(4):306–314, 2007. doi: 10.1107/S0108767307021903. URL `https://doi.org/10.1107/S0108767307021903`.

Emmanuel J. Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Found. Comput. Math.*, 14(5):1017–1026, October 2014. ISSN 1615-3375. doi: 10.1007/s10208-013-9162-z. URL `https://doi.org/10.1007/s10208-013-9162-z`.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, April 2015. ISSN 1557-9654. doi: 10.1109/TIT.2015.2399924. URL `https://ieeexplore.ieee.org/document/7029630/?arnumber=7029630`.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Ching-Hsien Chen and Jwo-Yuh Wu. Amplitude-aided 1-bit compressive sensing over noisy wireless sensor networks. *IEEE Wireless Communications Letters*, 4(5):473–476, 2015.

Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 822–883. Curran Associates, Inc., 2015. URL `https://papers.nips.cc/paper_files/paper/2015/hash/7380ad8a673226ae47fce7bff88e9c33-Abstract.html`.

Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020.

Aldo Conca, Dan Edidin, Milena Hering, and Cynthia Vinzant. An algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 38(2):346–356, 2015. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2014.

06.005. URL https://www.sciencedirect.com/science/article/pii/S1063520314000876.

J. V. Corbett. The pauli problem, state reconstruction and quantum-real numbers. *Reports on Mathematical Physics*, 57(1):53–68, 2006. ISSN 0034-4877. doi: 10.1016/S0034-4877(06)80008-X. URL https://www.sciencedirect.com/science/article/pii/S003448770680008X.

Aniket Das, Dheeraj M. Nagaraj, Praneeth Netrapalli, and Dheeraj Baby. Near optimal heteroscedastic regression with symbiotic learning. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3696–3757. PMLR, 12-15 Jul 2023. URL https://proceedings.mlr.press/v195/das23a.html.

Laurent Demanet and Vincent Jugnon. Convex recovery from interferometric measurements. *IEEE Transactions on Computational Imaging*, 3(2):282–295, 2017. doi: 10.1109/TCI.2017.2688923. URL https://ieeexplore.ieee.org/document/7895216.

Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 1 edition, August 2023a. ISBN 978-1-108-94316-1 978-1-108-83781-1. doi: 10.1017/9781108943161. URL https://www.cambridge.org/core/product/identifier/9781108943161/type/book. A draft is available from http://www.iliasdiakonikolas.org/ars-book.pdf.

Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, Cambridge, 2023b. ISBN 9781108837811. doi: 10.1017/9781108943161. URL https://www.cambridge.org/core/product/6AB511A1129F08F22D7D492B1A75ECAB.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. doi: 10.1137/17M1126680. URL https://doi.org/10.1137/17M1126680.

Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier Robust Mean Estimation with Subgaussian Rates via Stability. In *Advances in Neural Information Processing Systems*, volume 33, pages 1830–1840. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/13ec9935e17e00bed6ec8f06230e33a9-Abstract.html.

Ilias Diakonikolas, Daniel Kane, Jasper Lee, and Ankit Pensia. Outlier-robust sparse mean estimation for heavy-tailed distributions. *Advances in Neural Information Processing Systems*, 35:5164–5177, 2022.

Ilias Diakonikolas, Daniel M. Kane, Ankit Pensia, and Thanasis Pittas. Nearly-linear time and streaming algorithms for outlier-robust pca. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pages 7886–7921. JMLR.org, 2023. doi: 10.5555/3618408.3618721. URL https://dl.acm.org/doi/10.5555/3618408.3618721.

Haichen Dong, Alessio Mazzetto, Yu Cheng, and Rong Ge. Outlier-robust phase retrieval in nearly-linear time, 2024. URL https://openreview.net/forum?id=RTaSjoxmw9&referrer=%5Bthe%20profile%20of%20Rong%20Ge%5D(%2Fprofile%3Fid%3D~Rong_Ge1).

Haichen Dong, Alessio Mazzetto, Yu Cheng, and Rong Ge. Outlier-robust phase retrieval in nearly-linear time, 2025. URL https://openreview.net/forum?id=hEA2n6OUUK&referrer=%5Bthe%20profile%20of%20Rong%20Ge%5D(%2Fprofile%3Fid%3D~Rong_Ge1).

John Duchi, Saminul Haque, and Rohith Kuditipudi. A fast and slightly robust covariance estimator. *arXiv preprint arXiv:2502.20708*, 2025. doi: 10.48550/arXiv.2502.20708. URL https://arxiv.org/abs/2502.20708.

John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, September 2019. ISSN 2049-8772. doi: 10.1093/imaiai/iay015. URL https://doi.org/10.1093/imaiai/iay015.

C Fienup and J Dainty. Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application*, pages 231–275, 1987. URL https://www.researchgate.net/profile/James-Fienup/publication/247171131_Phase_retrieval_and_image_reconstruction_for_astronomy/links/594e7f2faca27248ae38564d/Phase-retrieval-and-image-reconstruction.pdf.

J. R. Fienup. Reconstruction of an object from the modulus of its fourier transform. *Opt. Lett.*, 3(1):27–29, Jul 1978. doi: 10.1364/OL.3.000027. URL https://opg.optica.org/ol/abstract.cfm?URI=ol-3-1-27.

J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, August 1982. doi: 10.1364/AO.21.002758. URL https://opg.optica.org/ao/abstract.cfm?URI=ao-21-15-2758.

N. C. Gallagher and B. Liu. Method for computing kinoforms that reduces image reconstruction error. *Applied Optics*, 12(10): 2328–2335, October 1973. doi: 10.1364/AO.12.002328. URL https://opg.optica.org/ao/abstract.cfm?URI=ao-12-10-2328.

R. W. Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35(2):237–246, 1972. URL https://api.semanticscholar.org/CorpusID:55691159.

Jean-Jacques Godeme, Jalal Fadili, Claude Amra, and Myriam Zerrad. Stable Phase Retrieval with Mirror Descent, June 2024. URL http://arxiv.org/abs/2405.10754. arXiv:2405.10754 [cs, math].

Paul Hand and Vladislav Voroninski. Corruption robust phase retrieval via linear programming. *arXiv preprint arXiv:1612.03547*, December 2016. doi: 10.48550/arXiv.1612.03547. URL https://arxiv.org/abs/1612.03547.

Robert W Harrison. Phase problem in crystallography. *Journal of the Optical Society of America A*, 10(5):1046–1055, May 1993. doi: 10.1364/JOSAA.10.001046. URL https://opg.optica.org/josaa/abstract.cfm?URI=josaa-10-5-1046.

Teiko Heinosaari, Luca Mazzarella, and Michael M. Wolf. Quantum tomography under prior information. *Communications in Mathematical Physics*, 318 (2):355–374, March 2013. ISSN 1432-0916. doi: 10.1007/s00220-013-1671-8. URL https://doi.org/10.1007/s00220-013-1671-8.

Peter M Hirsch, James A Jordan Jr, and Louis B Lesem. Method of making an object dependent diffuser, November 9 1971. URL https://patents.google.com/patent/US3619022A/en. US Patent 3,619,022.

Matthew J. Holland and Kazushi Ikeda. Efficient learning with robust gradient descent. *Mach. Learn.*, 108(8–9):1523–1560, September 2019. ISSN 0885-6125. doi: 10.1007/s10994-019-05802-5. URL https://doi.org/10.1007/s10994-019-05802-5.

Sam Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11902–11912. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc//paper_files/paper/2020/hash/8a1276c25f5efe85f0fc4020fbf5b4f8-Abstract.html.

Weihao Kong, Raghav Somani, Sham Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 4683–4696, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. URL https://dl.acm.org/doi/10.5555/3495724.3496117.

Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 1035–1046, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188970. URL https://doi.org/10.1145/3188745.3188970.

Jerry Li and Guanghao Ye. Robust gaussian covariance estimation in nearly-matrix multiplication time. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12649–12659. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9529fbba677729d3206b3b9073d1e9ca-Paper.pdf.

Shuyao Li, Yu Cheng, Ilias Diakonikolas, Jelena Diakonikolas, Rong Ge, and Stephen Wright. Robust second-order nonconvex optimization and its application to low rank matrix sensing. *Advances in Neural Information Processing Systems*, 36:54386–54398, 2023. URL https://neurips.cc/virtual/2023/poster/72944.

Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust m-estimation: Arbitrary corruption and heavy tails. *arXiv preprint arXiv:1901.08237*, 2019.

Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pages 411–421. PMLR, 2020.

Gábor Lugosi and Gergely Neu. Online-to-pac conversions: Generalization bounds via regret analysis. *arXiv:2305.19674*, 2023. URL https://arxiv.org/abs/2305.19674.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3345–3354. PMLR, 10-15 Jul 2018. URL https://proceedings.mlr.press/v80/ma18c.html.

Ibrahim Merad and Stéphane Gaïffas. Robust methods for high-dimensional linear learning. *Journal of Machine Learning Research*, 24(165):1–44, 2023.

Jing Miao, Tetsuya Ishikawa, Qun Shen, and Thomas Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annual Review of Physical Chemistry*, 59:387–410, 2008. doi: 10.1146/annurev.physchem.59.032607.093642. URL https://doi.org/10.1146/annurev.physchem.59.032607.093642.

R. P. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America A*, 7(3):394–411, March 1990. doi: 10.1364/JOSAA.7.000394. URL https://opg.optica.org/josaa/abstract.cfm?URI=josaa-7-3-394.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, September 2015. ISSN 1941-0476. doi: 10.1109/TSP.2015.2448516. URL https://ieeexplore.ieee.org/document/7130654/?arnumber=7130654.

Roberto I Oliveira and Zoraida F Rico. Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails. *The Annals of Statistics*, 52(5):1953–1977, 2024. URL https://projecteuclid.org/search?term=Improved+covariance+estimation%3A+optimal+robustness+and+sub-gaussian+guarantees+under+heavy+tails.

Soufiene Ben Othman, Abdelbasset Trad, Habib Youssef, and Hani Alzaid. Secure data aggregation in wireless sensor networks. In *2013 12th Annual Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET)*, pages 55–58. IEEE, 2013.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Journal of the American Statistical Association*, 120(550):1002–1013, 2025. doi: 10.1080/01621459.2024.2392906. URL https://doi.org/10.1080/01621459.2024.2392906.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):601–627, 04 2020. ISSN 1369-7412. doi: 10.1111/rssb.12364. URL https://doi.org/10.1111/rssb.12364.

Cheng Qian, Xiao Fu, N. Sidiropoulos, Lei Huang, and Junhao Xie. Inexact alternating optimization for phase retrieval in the presence of outliers. *IEEE Transactions on Signal Processing*, 65:6069–6082, 2016. URL https://api.semanticscholar.org/CorpusID:16444711.

Juri Ranieri, Amina Chebira, Yue M. Lu, and Martin Vetterli. Phase retrieval for sparse signals: Uniqueness conditions. *ArXiv*, abs/1308.3058, 2013. doi: 10.48550/arXiv.1308.3058. URL https://api.semanticscholar.org/CorpusID:11624243.

Takeyuki Sasai. Robust and sparse estimation of linear regression coefficients with heavy-tailed noises and covariates. *arXiv preprint arXiv:2206.07594*, 2022.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, October 2018. ISSN 1615-3383. doi: 10.1007/s10208-017-9365-9. URL https://doi.org/10.1007/s10208-017-9365-9.

Zhuojun Tian, Zhaoyang Zhang, and Lajos Hanzo. Distributed multi-view sparse vector recovery. *IEEE Transactions on Signal Processing*, 71:1448–1463, 2023.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018. ISBN 9781108231596. doi: https://doi.org/10.1017/9781108231596. URL https://www.cambridge.org/core/books/highdimensional-probability/797C466DA29743D2C8213493BD2D2102.

David Wagner. Resilient aggregation in sensor networks. In *Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks*, pages 78–87, 2004.

Irène Waldspurger. Phase Retrieval With Random Gaussian Sensing Vectors by Alternating Projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, May 2018. ISSN 1557-9654. doi: 10.1109/TIT.2018.2800663. URL https://ieeexplore.ieee.org/abstract/document/8276601.

Adriaan Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, 1963. doi: 10.1080/713817747. URL https://doi.org/10.1080/713817747.

Gang Wang, Georgios B. Giannakis, and Yonina C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Trans. Inf. Theor.*, 64(2):773–794, February 2018. ISSN 0018-9448. doi: 10.1109/TIT.2017.2756858. URL https://doi.org/10.1109/TIT.2017.2756858.

Yong Wang, Garhan Attebury, and Byrav Ramamurthy. A survey of security issues in wireless sensor networks. *IEEE Communications Surveys & Tutorials*, 8(2):2–23, 2006. doi: 10.1109/COMST.2006.315852.

Fan Wu and Patrick Rebeschini. Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent. *Information and Inference: A Journal of the IMA*, 12(2):633–713, June 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaac024. URL https://doi.org/10.1093/imaiai/iaac024.

Ming-Hsun Yang. Robust distributed phase retrieval for multi-view compressive networked sensing with outliers. *IEEE Wireless Communications Letters*, 2025.

Huishuai Zhang, Yi Zhou, Yingbin Liang, and Yuejie Chi. A nonconvex approach for phase retrieval: reshaped wirtinger flow and incremental algorithms. *J. Mach. Learn. Res.*, 18(1):5164–5198, January 2017. ISSN 1532-4435. doi: 10.5555/3122009.3208022. URL https://dl.acm.org/doi/abs/10.5555/3122009.3208022.

Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Median-Truncated Nonconvex Approach for Phase Retrieval With Outliers. *IEEE Transactions on Information Theory*, 64(11):7287–7310, November 2018. ISSN 1557-9654. doi: 10.1109/TIT.2018.2847695. URL https://ieeexplore.ieee.org/document/8386800/?arnumber=8386800.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Not Applicable.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes.
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes.
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] Not Applicable.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Not Applicable.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Not Applicable.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Not Applicable.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Not Applicable.
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Not Applicable.
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] Not Applicable.
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable.

A OTHER RELATED WORK

Phase retrieval. The literature on phase retrieval is both vast and diverse. Articles Balan et al. (2006) and Conca et al. (2015) have shown that $m \geq 2n - 1$ or $m \geq 4n - 4$ measurements $\{(a_i, y_i)\}_{i=1}^m$ are sufficient for uniquely determining an n-dimensional real-valued or complex-valued vector x^*, respectively, in the noise less settings. In the noiseless and no corruption setting, a wide array of algorithms have been proposed, including semidefinite programming (SDP) methods (Candès and Li, 2014) and nonconvex methods such as Wirtinger Flow (Candès et al., 2015), trust-region methods (Sun et al., 2018), alternating projections (Waldspurger, 2018), and variants like reshaped Wirtinger Flow (Zhang et al., 2017), truncated amplitude flow (TAF) (Wang et al., 2018) and AltMinPhase with resampling (Netrapalli et al., 2015). All of the above algorithm has sample complexities $\widetilde{O}(n)$. Beyond the noiseless setting, different noise models have been explored. Wu and Rebeschini (2023) addressed signal-independent light-tailed noise and proposed a mirror descent algorithm for sparse phase retrieval with sample complexity $\tilde{O}(k^2)$, where $\|x^*\|_0 = k$. Several works also studied bounded noise: Zhang et al. (2017) and Wang et al. (2018) also considered $|\zeta|_\infty = O(\|x^*\|)$, whereas Godeme et al. (2024) considered the case $|\zeta|_\infty = O(\|x^*\|^2)$ and proposed an efficient algorithm (spectral initialization and mirror descent) with the sample complexity $\tilde{O}(n)$. Chen and Candes (2015) presented a general framework that includes bounded and sub-exponential signal-dependent noise, highlighting two notable cases: $|\zeta|_\infty = O(\|x^*\|^2)$ and Poisson noise where $y_i \sim \text{Poisson}(|\langle a_i, x^* \rangle|^2)$. Their truncated Wirtinger Flow (TWF) algorithm uses gradient descent on a truncated empirical log-likelihood, discarding outliers at each step. In a related direction, Das et al. (2023) studied a heteroscedastic regression model with multiplicative noise of the form $y_i^2 = \chi^2(1)(\langle a_i, x^* \rangle)^2$, where $\chi^2(1)$ is a mean-one sub-exponential variable.

Robust statistics. Robust parameter estimation under heavy-tailed noise and adversarial corruption is a central theme in robust statistics, aiming for reliable learning under real-world data challenges. Suppose a dataset $\{x_i \in \mathbb{R}^d\}_{i=1}^n$ is drawn i.i.d. from a distribution $\mathcal{D}_\theta$, and an ϵ-fraction of the samples is arbitrarily corrupted. The goal is to robustly estimate θ from these corrupted samples. Robust statistics aims to estimate the mean (first-order) and covariance (second-order) statistical quantities under heavy-tailed noise and adversarial corruption. While the Tukey median achieves optimal mean estimation, it is computationally intractable. Diakonikolas et al. (2019) introduced the first polynomial-time algorithm with near-optimal error, which was later followed by nearly linear-time methods proposed in Hopkins et al. (2020). For covariance estimation, Oliveira and Rico (2024) proposed a statistically optimal but computationally intractable estimator, with efficient algorithms known only for Gaussian distributions (Kothari et al., 2018; Li and Ye, 2020), and for general distributions under low contamination (Duchi et al., 2025). Designing efficient algorithms for covariance estimation with optimal guarantees for arbitrary heavy-tailed distributions and constant corruption remains a key open problem. For principal component analysis (PCA), Kong et al. (2020) presented a polynomial-time algorithm for bounded random vectors, achieving almost linear sample complexity and tolerance to a constant fraction of corruptions. On the other hand, Diakonikolas et al. (2023) developed a nearly linear-time algorithm under the assumption that the data satisfies a certain stability condition (see Definition 4.6 for details).

B CORRECT ERROR SCALING FOR PHASE RETRIEVAL

In this section, we formally prove that any algorithm solving phase retrieval must have a $\frac{1}{\|x^*\|}$ dependence on the estimation error. Before presenting the proof, we first provide some intuition for this result using dimensional analysis. Note that the error scale σ in phase retrieval is $\mathcal{O}(\|x^*\|^2)$, while $dist(x, x^*) = \mathcal{O}(\|x^*\|)$. This implies a $\frac{1}{\|x^*\|}$ dependence by dimensional analysis. We now proceed to prove this statement formally.

Claim B.1. *Any algorithm that solves the phase retrieval problem (Model 1.2) with an estimation error of order $O(\sigma\sqrt{\epsilon}\|x^*\|^\alpha)$ must necessarily have $\alpha = -1$; in other words, the estimation error must scale inversely with the signal norm, i.e., as $1/\|x^*\|$.*

Proof. To address the above claim, we consider the following four exhaustive cases and show that each of them leads to a contradiction:

1. **Case 1.** $\text{dist}(x, x^*) = \mathcal{O}(\sigma\sqrt{\epsilon})$.
2. **Case 2.** $\text{dist}(x, x^*) = \mathcal{O}(\sigma\sqrt{\epsilon}\|x^*\|^\alpha)$, where $\alpha > 0$.
3. **Case 3.** $\text{dist}(x, x^*) = \mathcal{O}\left(\frac{\sigma\sqrt{\epsilon}}{\|x^*\|^\alpha}\right)$, where $0 < \alpha < 1$.

4. **Case 4.** $\mathrm{dist}(x, x^*) = \mathcal{O}\left(\frac{\sigma\sqrt{\epsilon}}{\|x^*\|^\alpha}\right)$, where $\alpha > 1$.

Case 1. Suppose there exists a phase retrieval algorithm whose estimation error does not depend on the signal norm $\|x^*\|$. Specifically, the algorithm outputs an estimate x_{out} such that

$$dist(x_{\text{out}}, x^*) = O\left(\sigma\sqrt{\epsilon}\right), \tag{7}$$

where σ is the noise level, ϵ is the fraction of corrupted samples. Assume $\|x^*\| = 1$, and there exists a dataset $S = (a_i, y_i)$ for this signal such that when the dataset S is given to the algorithm, it outputs x_{out} satisfying Equation (7).

Now fix some $n \in \mathbb{N}$, and consider the rescaled signal x^*/n. The dataset S can be converted into a dataset for this new signal simply by dividing each y_i by n^2. In this case, the measurement model becomes

$$\frac{y}{n^2} = \langle a, x^*/n\rangle^2 + \frac{\zeta}{n^2}.$$

Applying the algorithm to this rescaled dataset yields an output $\hat{x}_n^*$ such that

$$dist(\hat{x}_n^*, x^*/n) = O\left(\frac{\sigma}{n^2}\sqrt{\epsilon}\right),$$

which further implies

$$dist(n\hat{x}_n^*, x^*) = O\left(\frac{\sigma}{n}\sqrt{\epsilon}\right).$$

This shows that if there exists a algorithm for robust phase retrieval with absolute error independent of $\|x^*\|$ then it could be rescaled to solve the problem in the normalized setting $\|x^*\| = 1$ to arbitrary precision. This contradicts the known information-theoretic lower bound of $\mathcal{O}(\sigma\epsilon)$ for Gaussian linear regression (Bakshi and Prasad, 2021), which is a strictly easier problem than phase retrieval.

Case 2. Suppose there exists a phase retrieval algorithm whose estimation error follows case 2. Specifically, the algorithm outputs an estimate x_{out} such that

$$dist(x_{\text{out}}, x^*) = O\left(\sigma\sqrt{\epsilon}\|x^*\|^\alpha\right), \tag{8}$$

where $\alpha > 0$. Assume $\|x^*\| = 1$, and there exists a dataset $S = (a_i, y_i)$ for this signal such that when the dataset S is given to the algorithm, it outputs x_{out} satisfying Equation (8).

Now fix some $n \in \mathbb{N}$, and consider the rescaled signal x^*/n. The dataset S can be converted into a dataset for this new signal simply by dividing each y_i by n^2. In this case, the measurement model becomes

$$\frac{y}{n^2} = \langle a, x^*/n\rangle^2 + \frac{\zeta}{n^2}.$$

Applying the algorithm to this rescaled dataset yields an output $\hat{x}_n^*$ such that

$$dist(\hat{x}_n^*, x^*/n) = O\left(\frac{\sigma}{n^2}\sqrt{\epsilon}\frac{\|x^*\|^\alpha}{n^\alpha}\right) = O(\frac{\sigma}{n^{2+\alpha}}\sqrt{\epsilon}),$$

which further implies

$$dist(n\hat{x}_n^*, x^*) = O\left(\frac{\sigma}{n^{1+\alpha}}\sqrt{\epsilon}\right).$$

This implies that if there exists an algorithm for robust phase retrieval whose estimation error satisfies Equation (8), then the algorithm can be appropriately rescaled to solve the problem in the normalized setting $\|x^*\| = 1$ with arbitrarily high precision. Thus, a contradiction arises using the same argument as in case 1.

Case 3. Suppose there exists a phase retrieval algorithm whose estimation error follows case 3. Specifically, the algorithm outputs an estimate x_{out} such that

$$dist(x_{\text{out}}, x^*) = O\left(\sigma\sqrt{\epsilon}/\|x^*\|^\alpha\right), \tag{9}$$

where $0 < \alpha < 1$. Assume $\|x^*\| = 1$, and there exists a dataset $S = (a_i, y_i)$ for this signal such that when the dataset S is given to the algorithm, it outputs x_{out} satisfying Equation (9).

Now fix some $n \in \mathbb{N}$, and consider the rescaled signal x^*/n. The dataset S can be converted into a dataset for this new signal simply by dividing each y_i by n^2. In this case, the measurement model becomes

$$\frac{y}{n^2} = \langle a, x^*/n \rangle^2 + \frac{\zeta}{n^2}.$$

Applying the algorithm to this rescaled dataset yields an output $\hat{x}_n^*$ such that

$$dist(\hat{x}_n^*, x^*/n) = O\left(\frac{\frac{\sigma}{n^2}\sqrt{\epsilon}}{\frac{\|x^*\|^\alpha}{n^\alpha}}\right) = O(\frac{\sigma}{n^{2-\alpha}}\sqrt{\epsilon}),$$

which further implies

$$dist(n\hat{x}_n^*, x^*) = O\left(\frac{\sigma}{n^{1-\alpha}}\sqrt{\epsilon}\right).$$

This implies that if there exists an algorithm for robust phase retrieval whose estimation error satisfies Equation (9), then the algorithm can be appropriately rescaled to solve the problem in the normalized setting $\|x^*\| = 1$ with arbitrarily high precision. Thus, a contradiction arises using the same argument as in case 1.

Case 4. Suppose there exists a phase retrieval algorithm whose estimation error follows case 4. Specifically, the algorithm outputs an estimate x_{out} such that

$$dist(x_{\text{out}}, x^*) = O\left(\sigma\sqrt{\epsilon}/\|x^*\|^\alpha\right), \tag{10}$$

where $\alpha > 1$. Assume $\|x^*\| = 1$, and there exists a dataset $S = (a_i, y_i)$ for this signal such that when the dataset S is given to the algorithm, it outputs x_{out} satisfying Equation (10).

Now fix some $n \in \mathbb{N}$, and consider the rescaled signal nx^*. The dataset S can be converted into a dataset for this new signal simply by multiplying each y_i by n^2. In this case, the measurement model becomes

$$yn^2 = \langle a, x^* n \rangle^2 + \zeta n^2.$$

Applying the algorithm to this rescaled dataset yields an output $\hat{x}_n^*$ such that

$$dist(\hat{x}_n^*, x^* n) = O\left(\frac{n^2\sigma\sqrt{\epsilon}}{\|x^*\|^\alpha n^\alpha}\right) = O(\sigma n^{2-\alpha}\sqrt{\epsilon}),$$

which further implies

$$dist(\hat{x}_n^*/n, x^*) = O\left(\sigma n^{1-\alpha}\sqrt{\epsilon}\right) = O\left(\frac{\sigma\sqrt{\epsilon}}{n^{\alpha-1}}\right).$$

This implies that if there exists an algorithm for robust phase retrieval whose estimation error satisfies Equation (10), then the algorithm can be appropriately rescaled to solve the problem in the normalized setting $\|x^*\| = 1$ with arbitrarily high precision. Thus, a contradiction arises using the same argument as in case 1.

□

C COMMENTS ON USING FRESH SAMPLES AT EACH ITERATION OF GRADIENT DESCENT

Remark 4.3 informally states that using a constant number of batches of fresh samples is not fundamentally different from using a single batch. In particular, if we collect $O(Tm')$ samples with $T = O(1)$ and $log(T)/\varepsilon^2 \leq m'$, and uniformly divide them (without replacement) into T batches, then each batch will contain at least a $(1+c)\epsilon$ fraction of corrupted samples, where c is a very small positive constant. We formally state the above claim below and provide a proof.

Claim C.1. *Let the total dataset consist of $N = CTm'$ samples for some absolute constant $C > 0$, out of which exactly $K = \varepsilon N$ samples are corrupted. We construct T subsets (or batches) by partitioning the N samples uniformly at random, that is, by drawing Cm' samples for each subset uniformly* without replacement *from the dataset. Let X_j denote the number of corrupted samples in the j-th subset. If $m' \geq C_1 \frac{\log(T)+\log(1/\delta)}{\epsilon^2}$, then, with probability at least $1-\delta$, every subset contains at most a 2ε fraction of corrupted samples, where $C_1 = \frac{1}{2C}$.*

Proof. Let X_j be the number of corrupted samples in subset j. So, $\mathbb{P}(X_j = k) = \frac{{}^K C_k \, {}^{N-K}C_{Cm'-k}}{{}^N C_{Cm'}}$. This implies X_j follows a Hypergeometric distribution with parameter (N, K, Cm'). So, $\mathbb{E}[X_j] = \frac{Cm'K}{N} = \epsilon Cm'$.

Tail bound (Hypergeometric). Let $X \sim \text{Hypergeometric}(N, K, n')$ and set $p = K/N$. For any t with $0 < t < p$, the following tail bounds hold:

$$\mathbb{P}\big[X \leq (p-t)n'\big] \leq \exp(-n' \, D(p-t \,\|\, p)),$$
$$\mathbb{P}\big[X \geq (p+t)n'\big] \leq \exp(-n' \, D(p+t \,\|\, p)),$$

where the binary Kullback–Leibler divergence is $D(q \,\|\, p) \;=\; q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}$. Moreover, since $D(p \pm t \,\|\, p) \geq 2t^2$, one obtains the simpler Hoeffding-type bounds:

$$\mathbb{P}\big[X \leq (p-t)n'\big] \leq e^{-2t^2 n'},$$
$$\mathbb{P}\big[X \geq (p+t)n'\big] \leq e^{-2t^2 n'}.$$

So, applying the above tail bound for the hypergeometric distribution in our case implies for $0 \leq t \leq \epsilon$,

$$\mathbb{P}[X_j \geq (\epsilon + t)Cm'] \leq e^{-2Ct^2 m'}.$$

Union bound over the T subsets. Applying the tail bound to each subset and union bounding,

$$\mathbb{P}\big[\exists j \in [T] \;:\; X_j \geq (\epsilon + t)Cm'\big] \leq T \cdot e^{-2Ct^2 m'}.$$

Thus with probability at least $1 - T \cdot e^{-2Ct^2 m'}$, every subset satisfies

$$\frac{X_j}{Cm'} \leq (\epsilon + t).$$

Choose δ for a desired failure probability. Fix a target failure probability $\delta \in (0,1)$. To ensure the RHS $\leq \delta$, pick δ solving

$$T \cdot e^{-2Ct^2 m'} \leq \delta \quad \Longrightarrow \quad t \geq \sqrt{\frac{(\log T + \log(1/\delta))}{2Cm'}}.$$

Hence, with probability at least $1-\delta$, every subset has corrupted fraction

$$\frac{X_j}{Cm'} \leq \left(\varepsilon + \sqrt{\frac{(\log T + \log(1/\delta))}{2Cm'}} \right) = 2\epsilon.$$

□

D ADDITIONAL GUARANTEES OF ROBUST PCA ALGORITHM

D.1 Proof of Equation (4)

As the Remark 3.3 suggests that Algorithm 2 of Kong et al. (2020) has been designed to retrieve the top eigenvector for the case $k = 1$. Here , we rigorously proof Equation (4). Before proving Equation (4), we state a lemma from Kong et al. (2020) that will be used in the proof. The lemma is as follows:

Lemma D.1 (Main Lemma (Lemma C.4) for Algorithm 2 from (Kong et al., 2020))**.** *Consider the settings given in Theorem 3.2. The Algorithm 2 from (Kong et al., 2020) outputs a dataset $S' \subseteq T$ satisfying the following for $\hat{M} = \frac{1}{|S'|} \sum_{X_i \in S'} X_i X_i^\top$:*

1. *For the top normalized eigenvector $\hat{u} \in \mathbb{R}^{n\times 1}$ of $\hat{M}$,*

$$\operatorname{Tr}\left[\hat{u}^\top\left(\hat{M}-\Sigma_r\right)\hat{u}\right] \le 48\epsilon \operatorname{Tr}\left[\hat{u}^\top\Sigma_r\hat{u}\right] + 102v'\sqrt{\epsilon}.$$

2. *For all unit-norm vectors $V \in \mathbb{R}^{n\times 1}$, we have*

$$\operatorname{Tr}\left[V^\top\left(\hat{M}-\Sigma_r\right)V\right] \ge -10\epsilon \operatorname{Tr}[V^\top\Sigma_r V] - 8v'\sqrt{\epsilon}.$$

Now, we give a proof for Equation (4).

Theorem D.2. *In the settings of Theorem 3.2, if we also output the eigenvalue $\lambda_{\hat{u}}$ of the top eigenvector $\hat{u}$, then*

$$|\lambda_{\hat{u}} - \|\Sigma_r\|_{\mathrm{op}}| = O(\epsilon\|\Sigma_r\|_{\mathrm{op}} + \sqrt{\epsilon}v').$$

Proof. Since Σ_r is positive definite, the Eckart–Young–Mirsky theorem implies that

$$P_1(\Sigma_r) = \underset{B:\,\mathrm{rank}(B)=1}{\arg\min}\ \|\Sigma_r - B\|_{\mathrm{op}} = \lambda_1 v_1 v_1^\top,$$

where λ_1 and v_1 denote the largest eigenvalue and the corresponding normalized eigenvector of Σ_r, respectively. Consequently, $\operatorname{Tr}[P_1(\Sigma_r)] = \lambda_1 = \|\Sigma_r\|_{\mathrm{op}}$. . By using Lemma D.1, we can conclude that

$$|\lambda_{\hat{u}} - \hat{u}^\top\Sigma_r\hat{u}| \le 48\epsilon\hat{u}^\top\Sigma_r\hat{u} + 102\sqrt{\epsilon}v' \le 48\epsilon\|\Sigma_r\|_{\mathrm{op}} + 102\sqrt{\epsilon}v'. \tag{11}$$

And using the Theorem 3.2, we know that

$$\|\Sigma_r\|_{\mathrm{op}} - \hat{u}^\top\Sigma_r\hat{u} = \operatorname{Tr}[P_1(\Sigma_r)] - \hat{u}^\top\Sigma_r\hat{u} \le 58\epsilon\operatorname{Tr}[P_1(\Sigma_r)] + 110\sqrt{\epsilon}v' = 58\|\Sigma_r\|_{\mathrm{op}}\epsilon + 110\sqrt{\epsilon}v'. \tag{12}$$

Therefore, applying the triangle inequality together with Equation (11) and Equation (12), we obtain that

$$|\lambda_{\hat{u}} - \|\Sigma_r\|_{\mathrm{op}}| \le 106\epsilon\|\Sigma_r\|_{\mathrm{op}} + 212\sqrt{\epsilon}v' = O(\epsilon\|\Sigma_r\|_{\mathrm{op}} + \sqrt{\epsilon}v').$$

□

E ROBUST PHASE RETRIEVAL UNDER ZERO-MEAN NOISE

E.1 Proof of Lemma 4.4

Before proceeding with the proof, we state a special case of Lemma H.2, which will be used repeatedly. Specifically, for $i = 8$, Lemma H.2 gives that:

$$\mathbb{E}\left[\left(a^\top x^*\right)^8\left(v^\top a\right)^4\right] \le 10395\,\|x^*\|^4(v^\top x^*)^4 + \|x^*\|^8(315+5670)\,\|v\|^4 \le 16380\,\|x^*\|^8\|v\|^4 = O(\|x^*\|^8\|v\|^4), \tag{13}$$

for any fixed vector $v \in \mathbb{R}^{n\times 1}$. We now present the main proof.

Proof. It will be useful to bound the tail probability $p = \mathbb{P}(\|X\| \ge \tau)$:

$$\begin{aligned} p = \mathbb{P}(\|X\|\ge\tau) &\le \mathbb{E}[\|X\|^4]/\tau^4 = \mathbb{E}[(\langle a,x^*\rangle^2+\zeta)^4\|a\|^4]/\tau^4 \le \mathbb{E}\left[8n(\langle a,x^*\rangle^8+\zeta^4)\left(\sum_{j=1}^n a_j^4\right)/\tau^4\right] \\ &\overset{(a)}{\le} \left(8n^2(16380\|x^*\|^8) + 8nK_4^4\mathbb{E}\left[\left(\sum_{j=1}^n a_j^4\right)\right]\right)/\tau^4 = O\left(\frac{n^2(\|x^*\|^8+K_4^4)}{\tau^4}\right), \end{aligned} \tag{14}$$

where (a) follows from Equation (13). Decompose Σ as

$$\Sigma = \mathbb{E}[XX^\top] = \mathbb{E}[XX^\top\mathbb{I}_{\|X\|\le\tau}] + \mathbb{E}[XX^\top\mathbb{I}_{\|X\|\ge\tau}] = \mathbb{E}[\widetilde{X}\widetilde{X}^\top] + \mathbb{E}[XX^\top\mathbb{I}_{\|X\|\ge\tau}] = \widetilde{\Sigma}_r + \mathbb{E}[XX^\top\mathbb{I}_{\|X\|\ge\tau}]. \tag{15}$$

Upper bounding

$$\|\mathbb{E}[XX^\top \mathbb{I}_{\|X\|\geq\tau}]\|_{\text{op}} = \sup_{\|v\|=1} \mathbb{E}[\langle v, X\rangle^2 \mathbb{I}_{\|X\|\geq\tau}] \overset{(a)}{\leq} \sup_{\|v\|=1} \sqrt{p}\sqrt{\mathbb{E}[\langle v, X\rangle^4]} \leq \sup_{\|v\|=1} \sqrt{p}\sqrt{\mathbb{E}[8(\zeta^4 + \langle a, x^*\rangle^8)\langle a, v\rangle^4]}$$
$$\overset{(b)}{\leq} \sqrt{p}\sqrt{8(3K_4^4 + 16380\|x^*\|^8)} = O\left(\sqrt{p}\sqrt{(K_4^4 + \|x^*\|^8)}\right) = O(\sqrt{p}(K_4^2 + \|x^*\|^4)),$$

where (a) follows from Cauchy-Schwarz inequality and (b) follows from Equation (13) and using that $\sqrt{p} = O(n(K_4^2 + \|x^*\|^4)/\tau^2)$ (by Equation (14)), we have

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\text{op}} = \mathbb{E}[XX^\top \mathbb{I}_{\|X\|\geq\tau}]\|_{\text{op}} = O(n(K_4^2 + \|x^*\|^4)^2/\tau^2).$$

□

E.2 Proof of the Theorem 4.5

Before presenting the proof of Theorem 4.5, we introduce two helper theorems. These theorems provide the guarantees for the first two steps of Algorithm 1. We first analyze the step 1 of Algorithm 1

Theorem E.1. ***(Step 1 of Algorithm 1: Estimate $\|x^*\|$ using a robust mean estimator).*** *(See also the part "scaling" from the proof of Theorem 3.1 in Buna and Rebeschini (2025).) If $m_1 = \Omega\left(\log(2/\delta)\left(1 + r_{up}^2\right)\right)$ and $\epsilon = O\left(\left(1 + r_{up}^2\right)^{-1}\right)$, then with probability at least $1 - \delta/2$ we can conclude that*

$$|\|\hat{x^*}\| - \|x^*\|| \leq \frac{1}{27}\|x^*\|. \tag{16}$$

Proof. Note that $\mathbb{E}[y] = \mathbb{E}\left[\left(a^\top x^*\right)^2 + \zeta\right] = \|x^*\|^2$ and $\text{Var}(y) = \mathbb{E}\left[\left(a^\top x^*\right)^4\right] + \mathbb{E}\left[(\zeta)^2 \mid a\right] - \|x^*\|^4 = 2\|x^*\|^4 + \sigma^2$. Then, using Theorem 3.1 we can say that, with probability at least $1 - \delta/2$:

$$\left|\|\hat{x^*}\|^2 - \|x^*\|^2\right| = O\left(\sqrt{\text{Var}(y)}\left(\sqrt{\epsilon} + \sqrt{\frac{\log(2/\delta)}{m_1}}\right)\right) = O\left(\sqrt{\|x^*\|^4 + \sigma^2}\left(\sqrt{\epsilon} + \sqrt{\frac{\log(2/\delta)}{m_1}}\right)\right)$$

By using Cauchy-Schwarz inequality, we can say that $\sigma^2 \leq K_4^2$, which with the choices of m_1 and ϵ concludes that $\sqrt{\left(1 + \frac{\sigma^2}{\|x^*\|^4}\right)}\left(\sqrt{\epsilon} + \sqrt{\frac{\log(2/\delta)}{m_1}}\right) \leq \sqrt{\left(1 + r_{up}^2\right)}\left(\sqrt{\epsilon} + \sqrt{\frac{\log(2/\delta)}{m_1}}\right) \leq 1$. Now, if $\widetilde{y}^2 > 0$, then

$$\left|\|\hat{x^*}\|^2 - \|x^*\|^2\right| = O\left(\|x^*\|^2\sqrt{\left(1 + \frac{\sigma^2}{\|x^*\|^4}\right)}\left(\sqrt{\epsilon} + \sqrt{\frac{\log(2/\delta)}{m_1}}\right)\right).$$

Now, we know that for any $a, b, c > 0$ with $b^2 > c$, $\left|a^2 - b^2\right| \leq c$ implies $|a - b| \leq b - \sqrt{b^2 - c}$. By using the above property, we get to

$$|\|\hat{x^*}\| - \|x^*\|| = O\left(\|x^*\|\left(1 - \sqrt{1 - \sqrt{\left(1 + \frac{\sigma^2}{\|x^*\|^4}\right)}\left(\sqrt{\epsilon} + \sqrt{\frac{\log(2/\delta)}{m_1}}\right)}\right)\right).$$

Now, using the bound $\left(1 + \frac{\sigma^2}{\|x^*\|^4}\right) \leq \left(1 + r_{up}^2\right)$, we can conclude that

$$\left|\|\hat{x^*}\| - \|x^*\|\right| = O\left(\|x^*\|\left(1 - \sqrt{1 - \sqrt{1 + r_{up}^2}\left(\sqrt{\epsilon} + \sqrt{\frac{\log(2/\delta)}{m_1}}\right)}\right)\right).$$

Now, we can choose the hidden in such a way that

$$||\hat{\|x^*\|} - \|x^*\|\,| \leq \frac{1}{27}\,\|x^*\|\,. \tag{17}$$

□

Next, we analyze the step 2 of Algorithm 1.

Theorem E.2. *(**Step 2 of Algorithm 1: Estimating** $\hat{u}$.) Let's assume that* $||\hat{\|x^*\|} - \|x^*\|\,| \leq \frac{\|x^*\|}{27}$ *and* $\delta \in (0, 0.5)$ *is a positive small constant. If* $m_2 = \Omega\left((n + n(r_{up}^2+1)\sqrt{\epsilon})\log(2n/(\delta\epsilon))/\epsilon\right)$ *and* $\epsilon \in (0, 1/36]$*. Then, under the above notation, with probability at least* $1 - \delta/2$*, the following holds:*

$$\mathrm{dist}(\hat{u}, u) = \min\{\|\hat{u} - u\|_2, \|\hat{u} + u\|_2\} \leq \sqrt{2 - 2\sqrt{1 - O\left((r_{up}^2+1)\sqrt{\epsilon}\right)}},$$

where u *is the top normalized eigenvector of* $\widetilde{\Sigma}_r$*.*

Proof. We use Theorem 3.2 to calculate the distance between $\hat{u}$ and the top eigenvector u of $\widetilde{\Sigma}_r$. To apply Theorem 3.2, we must verify that all its underlying assumptions are satisfied. We now proceed to check these assumptions one by one. We know that $\|\widetilde{X}\| \leq \hat{\tau}$. Lemma 4.4 implies

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\mathrm{op}} = O(\,n(K_4^2 + \|x^*\|^4)^2/\hat{\tau}^2) \leq \|x^*\|^4 \quad \text{and} \quad -\|x^*\|^4 I \preceq \widetilde{\Sigma}_r - \Sigma \preceq \|x^*\|^4 I.$$

The first assumption is to show that $\|\widetilde{X}\widetilde{X}^\top - \widetilde{\Sigma}_r\|_{\mathrm{op}} \leq B$ for all $\widetilde{X}$ with probability one.

$$\begin{aligned}\|\widetilde{X}\widetilde{X}^\top - (\widetilde{\Sigma}_r)\|_{\mathrm{op}} &\leq \|\widetilde{X}\|^2 + \lambda_{max}(\widetilde{\Sigma}_r) \leq \hat{\tau}^2 + \lambda_{max}(\Sigma) + \|x^*\|^4 = O(n(r_{up}^2+1)^2\hat{x}^{*4}) + \left(16\|x^*\|^4 + \sigma^2\right) \\ &= O\left(n(r_{up}^2+1)^2 + 16\right)\|x^*\|^4 + \sigma^2.\end{aligned}$$

The second assumption is to show that $\mathbb{E}\left[\langle xx^\top, \widetilde{X}\widetilde{X}^\top - \widetilde{\Sigma}_r\rangle^2\right] \leq v'^2$.

Consider $\|x\| \leq 1$, then

$$\begin{aligned}\mathbb{E}\left[\langle xx^\top, \widetilde{X}\widetilde{X}^\top - \widetilde{\Sigma}_r\rangle^2\right] &= \mathbb{E}\left[(\langle x, \widetilde{X}\rangle^2 - x^\top\widetilde{\Sigma}_r x)^2\right] = \mathbb{E}\left[(\langle x, \widetilde{X}\rangle)^4\right] - (x^\top\widetilde{\Sigma}_r x)^4 \overset{(a)}{\leq} \mathbb{E}\left[(\langle x, X\rangle)^4\right] \\ &\leq \mathbb{E}[8(\zeta^4 + \langle a, x^*\rangle^8)\langle a, x\rangle^4] \overset{(b)}{\leq} 8(3K_4^4 + 16380\|x^*\|^8) = O((r_{up}^2+1)^2\|x^*\|^8),\end{aligned}$$

where (a) follows by applying the trivial upper bound $\hat{\tau} \leq \infty$ and (b) follows from Equation (13).

This implies

$$\max_{x:\|x\|\leq 1} \mathbb{E}\left[\langle xx^\top, \widetilde{X}\widetilde{X}^\top - \widetilde{\Sigma}\rangle^2\right] = O((r_{up}^2+1)^2\|x^*\|^8).$$

In the notations of Theorem 3.2, we have

$$B = O\left(\,n\,(r_{\mathrm{up}}^2+1)^2 + 16\right)\|x^*\|^4 + \sigma^2, \quad v' = O(\,(r_{\mathrm{up}}^2+1)\,\|x^*\|^4).$$

To apply Theorem 3.2, we must ensure that its sample complexity condition is satisfied. This requires

$$m_2 = \Omega\left(\frac{n + \frac{B\sqrt{\epsilon}}{v'}}{\epsilon} \cdot \log\left(\frac{2n}{\delta\epsilon}\right)\right) = \Omega\left(\frac{n + \,n\,(r_{\mathrm{up}}^2+1)\sqrt{\epsilon}}{\epsilon} \cdot \log\left(\frac{2n}{\delta\epsilon}\right)\right).$$

So, we have verified all assumptions of Theorem 3.2. So, by using Theorem 3.2, we can conclude that, with probability at least $1 - \delta/2$, the output $\hat{u} \in \mathbb{R}^{n\times 1}$ of Algorithm 2 satisfies:

$$\mathrm{Tr}[P_1(\widetilde{\Sigma}_r)] - \mathrm{Tr}[\hat{u}^\top(\widetilde{\Sigma}_r)\hat{u}] = O\left(\epsilon \cdot \mathrm{Tr}[P_1(\widetilde{\Sigma}_r)] + \sqrt{\epsilon}\cdot v'\right).$$

We know that the top eigenvalue of Σ is $15\|x^*\|^4 + \sigma^2$, while all remaining eigenvalues equal $3\|x^*\|^4 + \sigma^2$. Let $\widetilde{\Sigma}_r = \sum_{i=1}^n \lambda_i f_i f_i^\top$ be the eigenvalue decomposition of $\widetilde{\Sigma}_r$, with $\lambda_1 = \lambda_{\max}(\widetilde{\Sigma}_r)$ and $f_1 = u$ such that eigenvalues of $\widetilde{\Sigma}_r$ are arranged in non-increasing order, where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A. By the min–max theorem, we have $\lambda_1 \in [14\|x^*\|^4 + \sigma^2,\ 16\|x^*\|^4 + \sigma^2]$ and $\lambda_2 \in [2\|x^*\|^4 + \sigma^2,\ 4\|x^*\|^4 + \sigma^2]$. Since $4\|x^*\|^4 + \sigma^2 \neq 14\|x^*\|^4 + \sigma^2$, it follows that $\lambda_1 \neq \lambda_2$.

$$\begin{aligned}\mathrm{Tr}[P_1(\widetilde{\Sigma}_r)] - \mathrm{Tr}[\hat{u}^\top(\widetilde{\Sigma}_r)\hat{u}] &\geq \lambda_1 - (\lambda_1\langle \hat{u}, u\rangle^2 + (1 - \langle \hat{u}, u\rangle^2)\lambda_2) = (\lambda_1 - \lambda_2)(1 - \langle \hat{u}, u\rangle^2) \\ &\geq 10\|x^*\|^4(1 - \langle \hat{u}, u\rangle^2).\end{aligned}$$

So now,

$$(1 - \langle \hat{u}, u\rangle^2) = O\left(\epsilon\left(1 + \frac{\sigma^2}{\|x^*\|^4}\right) + 73(r_{up}^2 + 1)\sqrt{\epsilon}\right) = O\left(\epsilon(r_{up}^2 + 1) + (r_{up}^2 + 1)\sqrt{\epsilon}\right) = O\left((r_{up}^2 + 1)\sqrt{\epsilon}\right).$$

This implies

$$|\langle \hat{u}, u\rangle| \geq \sqrt{1 - O\left((r_{up}^2 + 1)\sqrt{\epsilon}\right)}.$$

Now,

$$\begin{aligned}\mathrm{dist}(\hat{u}, u) &= \min\{\|\hat{u} - u\|_2, \|\hat{u} + u\|_2\} \\ &= \sqrt{2 - 2|\langle \hat{u}, u\rangle|} \\ &\leq \sqrt{2 - 2\sqrt{1 - O\left((r_{up}^2 + 1)\sqrt{\epsilon}\right)}}.\end{aligned}$$

□

We are now in a position to prove Theorem 4.5.

Proof of Theorem 4.5.

Proof. Under the assumptions on the sample complexity m_1, m_2 and the corruption level ϵ, and using Theorem E.1, Theorem E.2, and Lemma 4.4, we conclude that, with probability at least $1 - \delta$, the following holds:

$$\mathrm{dist}(\hat{u}, u) = \min\{\|\hat{u} - u\|_2, \|\hat{u} + u\|_2\} \leq \sqrt{2 - 2\sqrt{1 - O\left((r_{up}^2 + 1)\sqrt{\epsilon}\right)}} \leq 1/27$$

,

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\mathrm{op}} = O(n(K_4^2 + \|x^*\|^4)^2/\hat{\tau}^2) \leq \frac{\sqrt{2}\|x^*\|^4}{9},$$
$$|\|\hat{x^*}\| - \|x^*\|| \leq \|x^*\|/27,$$

where u is top normalized eigenvector of $\widetilde{\Sigma}_r$. Now, by applying Theorem H.3 (Davis-Kahan), we can conclude that

$$\begin{aligned}dist(\|x^*\|u, x^*) &\leq \frac{2\sqrt{2}\|x^*\|\|\Sigma - (\widetilde{\Sigma}_r)\|_{\mathrm{op}}}{12\|x^*\|^4} \leq \frac{\|x^*\|}{27}, \quad \text{and} \\ dist(-\|x^*\|u, x^*) &\leq \frac{2\sqrt{2}\|x^*\|\|\Sigma - (\widetilde{\Sigma}_r)\|_{\mathrm{op}}}{12\|x^*\|^4} \leq \frac{\|x^*\|}{27}\end{aligned}$$

.

So,

$$\begin{aligned}\mathrm{dist}\,(x_0, x^*) = \mathrm{dist}\left(\|\hat{x^*}\|\hat{u}, x^*\right) &\leq \mathrm{dist}(\|\hat{x^*}\|\hat{u}, \|x^*\|\hat{u}) + \mathrm{dist}(\|x^*\|\hat{u}, \|x^*\|u) + dist\,(\|x^*\|u, x^*) \\ &= \|\|\hat{x^*}\|\hat{u} - \|x^*\|\hat{u}\| + \mathrm{dist}(\|x^*\|\hat{u}, \|x^*\|u) + dist\,(\|x^*\|u, x^*) \\ &= |\|\hat{x^*}\| - \|x^*\|| + \|x^*\|\,\mathrm{dist}(\hat{u}, u) + dist\,(\|x^*\|u, x^*) \\ &\leq \frac{\|x^*\|}{27} + \frac{\|x^*\|}{27} + \frac{\|x^*\|}{27} = \frac{\|x^*\|}{9}.\end{aligned}$$

□

F NEAR LINEAR TIME ALGORITHMS FOR ROBUST PHASE RETRIEVAL UNDER ZERO-MEAN NOISE

F.1 Guarantees of Algorithm 2 (SampleTopEigenvector) from (Diakonikolas et al., 2023)

Theorem F.1. *(See Theorem 3.1 from (Diakonikolas et al., 2023)) Let γ_0 be a sufficiently small positive constant. Let $n > 2$ be an integer, and $\epsilon, \gamma \in (0,1)$ such that $0 < 20\epsilon < \gamma < \gamma_0$. Let D be the uniform distribution over a set of m' points in $\mathbb{R}^n$, that can be decomposed as $D = (1-\epsilon)G + \epsilon B$, where G is a $(20\epsilon, \gamma)$-stable distribution (Definition 4.6) with respect to a PSD matrix $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$. Here, G denotes the set of uncorrupted samples and B denotes the set of corrupted samples. There exists an algorithm (Algorithm 2 from (Diakonikolas et al., 2023)) that takes as input D, ϵ, γ, runs for $O\left(\frac{m'n}{\gamma^2}\log^4(n/\epsilon)\right)$ time, and with probability at least 0.99 , outputs a unit vector $\hat{u}$ such that $\hat{u}^\top \mathbf{\Sigma} \hat{u} \geq (1 - O(\gamma))\|\mathbf{\Sigma}\|_{\mathrm{op}}$.*

F.2 Formal version of Proposition 4.7 and its proof

Here, we first present the formal version of Proposition 4.7 and then we prove it.

Proposition F.2. *Let's assume $\delta \in [0, 0.5]$ be a constant, $m_1 = \Omega(\log(2/\delta)(1 + r_{up}^2))$, $\tilde{m} = \Omega(n\log(n)\log(2/\delta)(1 + r_{up}^2))$ and $\epsilon \leq C_2/(1 + r_{up}^2)$. Let G be the empirical distribution over $f(n)$ uncorrupted samples of $\widetilde{X}$. If G is $(\epsilon, \gamma = 5/C_3 27^2(16 + r_{up}^2))$-stable with respect to $\widetilde{\Sigma}_r$, then there exists an algorithm for robust phase retrieval with the sample complexity $O(m_1 + f(n) + P\tilde{m})$ and time complexity $\widetilde{O}(T_{\text{rob-mean}}(m_1, 1) + f(n)n + PT_{\text{rob-mean}}(\tilde{m}, n))$ such that with probability at least $0.99 - (P+1)\delta$, the output x_{out} of the algorithm satisfies $dist(x_{out}, x^*) \leq \sigma\sqrt{\epsilon}/\|x^*\|$, where $P = O(1)$ is constant.*

Proof. The algorithm considered is A_{alt} + Algorithm 2 from (*BunaandRebeschini*, 2025), where A_{alt} is the same as Algorithm 1, except that in Step 2, Algorithm 2 of (Kong et al., 2020) is replaced with the `SampleTopEigenvector` (Algorithm 2) from (Diakonikolas et al., 2023). From the Theorem 4.2 it is evident that if the output x_0 of A_{alt} lies in $\mathcal{B}(\pm x^*, \|x^*\|/9)$, then the final output x_{out} of above algorithm satisfies $dist(x_{out}, x^*) \leq \sigma\sqrt{\epsilon}/\|x^*\|$ (*the algorithm successfully solves the phase retrieval problem*). So, we need to show that $x_0 \in \mathcal{B}(\pm x^*, \|x^*\|/9)$. Now, as the empirical distribution over $f(n)$ uncorrupted samples satisfies the stability lemma, using Theorem F.1, Theorem E.1, and Lemma 4.4, we conclude that, with probability at least $0.99 - \delta$, the following holds:

$$\|\widetilde{\Sigma}_r\|_{\mathrm{op}} - \hat{u}^\top \widetilde{\Sigma}_r \hat{u} \leq O(\gamma)\|\widetilde{\Sigma}_r\|_{\mathrm{op}}, \quad |\hat{\|x^*\|} - \|x^*\|| \leq \|x^*\|/27, \quad \|\Sigma - \widetilde{\Sigma}_r\|_{\mathrm{op}} = O(n(K_4^2 + \|x^*\|^4)^2/\hat{\tau}^2) \leq \frac{\sqrt{2}\|x^*\|^4}{9},$$

where u is top normalized eigenvector of $\widetilde{\Sigma}_r$ and $\hat{u}$ is the output of `SampleTopEigenvector` (Algorithm 2) from (Diakonikolas et al., 2023) upon providing $f(n)/(1-\epsilon)$ ($f(n)$ uncorrupted samples and $\epsilon f(n)/1-\epsilon$ corrupted samples) samples. Now, by applying Theorem H.3 (Davis-Kahan), we can conclude that

$$\mathrm{dist}(\|x^*\|u, x^*) \leq \frac{2\sqrt{2}\|x^*\|\|\Sigma - (\widetilde{\Sigma}_r)\|_{\mathrm{op}}}{12\|x^*\|^4} \leq \frac{\|x^*\|}{27} \quad \text{and}$$

$$\mathrm{dist}(-\|x^*\|u, x^*) \leq \frac{2\sqrt{2}\|x^*\|\|\Sigma - (\widetilde{\Sigma}_r)\|_{\mathrm{op}}}{12\|x^*\|^4} \leq \frac{\|x^*\|}{27}.$$

Note that $\|\Sigma - \widetilde{\Sigma}_r\|_{\mathrm{op}} \leq \|x^*\|^4$ implies $-\|x^*\|^4 \cdot I \preceq (\widetilde{\Sigma}_r) - \Sigma \preceq \|x^*\|^4 \cdot I$. Now, we know that the top eigenvalue of Σ is $15\|x^*\|^2 + \sigma^2$, and the rest of the eigenvalues are the same, with a value of $3\|x^*\|^4 + \sigma^2$. Assume that $\widetilde{\Sigma}_r$ has the eigenvalue decomposition $\widetilde{\Sigma}_r = \sum_{i=1}^n \lambda_i f_i f_i^\top$, with $\lambda_1 = \lambda_{\max}(\widetilde{\Sigma}_r)$, $f_1 = u$ such that eigenvalues of $\widetilde{\Sigma}_r$ are arranged in non-increasing order, where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A. Min-max theorem implies that $\lambda_1 \in [14\|x^*\|^4 + \sigma^2, 16\|x^*\|^4 + \sigma^2)]$ and $\lambda_2 \in [2\|x^*\|^4 + \sigma^2, 4\|x^*\|^4 + \sigma^2]$. Now, $4\|x^*\|^4 + \sigma^2 \neq 14\|x^*\|^4 + \sigma^2$ implies that $\lambda_1 \neq \lambda_2$.

$$\begin{aligned}
\|\widetilde{\Sigma}_r\|_{\mathrm{op}} - (\hat{u}^\top \widetilde{\Sigma}_r \hat{u}) &\geq \lambda_1 - (\lambda_1\langle \hat{u}, u\rangle^2 + (1 - \langle \hat{u}, u\rangle^2)\lambda_2) = (\lambda_1 - \lambda_2)(1 - \langle \hat{u}, u\rangle^2) \\
&\geq 10\|x^*\|^4(1 - \langle \hat{u}, u\rangle^2).
\end{aligned}$$

So, this implies

$$10\|x^*\|^4(1-\langle\hat{u},u\rangle^2) \le O(\gamma)(16\|x^*\|^4+\sigma^2) \le O(\gamma)(16+r_{up}^2)\|x^*\|^4 \le 5\|x^*\|^4/27^2$$

, which further concludes that

$$|\langle\hat{u},u\rangle| \ge \sqrt{1-\frac{1}{2\times 27^2}} \ge \left(1-\frac{1}{2\times 27^2}\right).$$

Now,

$$\begin{aligned}\mathrm{dist}(\hat{u},u) &= \min\{\|\hat{u}-u\|_2, \|\hat{u}+u\|_2\} \\ &= \sqrt{2-2|\langle\hat{u},u\rangle|} = \sqrt{2}\sqrt{1-|\langle\hat{u},u\rangle|} \le \sqrt{2}\sqrt{1-\left(1-\frac{1}{2\times 27^2}\right)} = \frac{1}{27}.\end{aligned}$$

So, from above we can conclude that $\mathrm{dist}(\hat{u},u) \le 1/27$. So the output x_0 of A_{alt} satisfies,

$$\begin{aligned}\mathrm{dist}\,(x_0,x^*) = \mathrm{dist}\left(\hat{\|x^*\|}\hat{u},x^*\right) &\le \mathrm{dist}(\hat{\|x^*\|}\hat{u},\|x^*\|\hat{u}) + \mathrm{dist}(\|x^*\|\hat{u},\|x^*\|u) + dist\,(\|x^*\|u,x^*) \\ &= \|\hat{\|x^*\|}\hat{u}-\|x^*\|\hat{u}\| + \mathrm{dist}(\|x^*\|\hat{u},\|x^*\|u) + dist\,(\|x^*\|u,x^*) \\ &= |\hat{\|x^*\|}-\|x^*\|| + \|x^*\|\,\mathrm{dist}(\hat{u},u) + dist\,(\|x^*\|u,x^*) \\ &\le \frac{\|x^*\|}{27}+\frac{\|x^*\|}{27}+\frac{\|x^*\|}{27} = \frac{\|x^*\|}{9}.\end{aligned}$$

The running time of the above algorithm is $\widetilde{O}(T_{\text{rob-mean}}(m_1,1)+f(n)n+PT_{\text{rob-mean}}(\tilde{m},n))$ and it follows from Theorem F.1. □

F.3 Lemma F.3 and its proof

Lemma F.3. *(Stability lemma). Assume that $|\hat{x}^*-\|x^*\|| \le \frac{1}{27}\|x^*\|$. Let $\delta\in[0,0.5]$, $p\ge 2$. Under the assumptions and notations from above if $m_2 \ge C_1\max\{n\log(4n/\delta)((16+r_{up}^2)^4), n^5(r_{up}^2+1)^4\log(4/\delta)\}$, and $\epsilon \le C_2(16+r_{up}^2)^{-4}$ where C_1, C_2 are universal constants. Then the following holds:*

$$\Pr\left[\exists w\in\mathfrak{S}_\epsilon^{m_2}\;\middle|\;\left\|\sum_i w_i\widetilde{X}_i\widetilde{X}_i^\top-\widetilde{\Sigma}_r\right\|_{\mathrm{op}} > \gamma\|\widetilde{\Sigma}_r\|_{\mathrm{op}}\right] \le \frac{\delta}{2}. \tag{18}$$

where $\mathfrak{S}_\epsilon^{m_2}=\{w\in\mathbb{R}^{m_2}:\sum_{i=1}^{m_2}w_i=1, w_i\le 1/(1-\epsilon)m_2\}$, and $\gamma<1$.

Before presenting the the proof of Lemma F.3, we first state a concentration lemma for the sample covariance matrix of bounded random vectors, which will be needed to establish the stability lemma.

Theorem F.4. ***(Matrix Bernstein inequality (See (Vershynin, 2018)).*** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^n$ supported within an ℓ_2-ball of radius τ centered at the origin. Let $\widetilde{\Sigma}_r$ denotes the second moment matrix of $\mathcal{D}$, and let $\widetilde{\Sigma}_{m'}=\frac{1}{m'}\sum_{i=1}^{m'}\widetilde{X}_i\widetilde{X}_i^\top$ be the empirical second moment matrix based on n i.i.d. samples $\widetilde{X}_1,\ldots,\widetilde{X}_{m'}\sim\mathcal{D}$. Then there exists a universal constant $C>0$ such that for any $0<t<1$ and $0<\delta<1$, if*

$$m' > C\cdot\frac{\tau^2}{t^2\|\widetilde{\Sigma}_r\|_{\mathrm{op}}}\log\left(\frac{n}{\delta}\right),$$

then with probability at least $1-\delta$, we have $\|\widetilde{\Sigma}_{m'}-\widetilde{\Sigma}_r\|_{\mathrm{op}} \le t\|\widetilde{\Sigma}_r\|_{\mathrm{op}}$.

Now we prove the Lemma F.3.

Proof. We prove the Lemma F.3 for $\gamma=\frac{5}{27^2(16+r_{up}^2)}$. Consider G to be a uniform distribution over a set of m_2 uncorrupted samples $\{\widetilde{X}_1,\widetilde{X}_2,\ldots,\widetilde{X}_{m_2}\}$. Then, for any weight function w, $\boldsymbol{\Sigma}_{G_w}=\sum_i w_i\widetilde{X}_i\widetilde{X}_i^\top$ for some $w\in\mathfrak{S}_\epsilon^{m_2}$. Let's denote the left-hand side of Equation (18) by $\widetilde{p}$, then

$$\widetilde{p} \le \Pr\left[\max_{w \in \mathfrak{S}_\epsilon^{m_2}} \left\| \sum_i w_i \widetilde{X}_i \widetilde{X}_i^\top - \widetilde{\Sigma}_r \right\|_{\text{op}} > 5/(27^2(16 + r_{up}^2)) \|\widetilde{\Sigma}_r\|_{\text{op}} \right].$$

Now for any $w \in \mathfrak{S}_\epsilon^{m_2}$,

$$\begin{aligned}
\left\| \sum_i w_i \widetilde{X}_i \widetilde{X}_i^\top - \widetilde{\Sigma}_r \right\|_{\text{op}} &= \left\| \sum_i w_i \widetilde{X}_i \widetilde{X}_i^\top - \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i \widetilde{X}_i^\top + \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i \widetilde{X}_i^\top - \widetilde{\Sigma}_r \right\|_{\text{op}} \\
&\le \left\| \sum_i w_i \widetilde{X}_i \widetilde{X}_i^\top - \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i \widetilde{X}_i^\top \right\|_{\text{op}} + \left\| \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i \widetilde{X}_i^\top - \widetilde{\Sigma}_r \right\|_{\text{op}}.
\end{aligned}$$

Now,

$$\begin{aligned}
\max_{w \in \mathfrak{S}_\epsilon^{m_2}} \left\| \sum_i w_i \widetilde{X}_i \widetilde{X}_i^\top - \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i \widetilde{X}_i^\top \right\|_{\text{op}} &= \max_{w \in \mathfrak{S}_\epsilon^{m_2}} \max_{v:\|v\|=1} \left| \sum_i \left(w_i - \frac{1}{m_2}\right) (\widetilde{X}_i^\top v)^2 \right| \\
&\le \max_{w \in \mathfrak{S}_\epsilon^{m_2}} \max_{v:\|v\|=1} \sqrt{\sum_i \left(w_i - \frac{1}{m_2}\right)^2} \sqrt{\sum_i \left(\widetilde{X}_i^\top v\right)^4} \\
&\le \sqrt{\frac{\epsilon}{(1-\epsilon)}} \max_{v:\|v\|=1} \sqrt{\frac{1}{m_2}\sum_i \left(\widetilde{X}_i^\top v\right)^4},
\end{aligned}$$

because $\sum_i (w_i - 1/m_2)^2 \le \sum_i w_i^2 - 1/m_2 \le \epsilon/((1-\epsilon)m_2))$. Now, note that for any v, $\left(\widetilde{X}_i^\top v\right)^4 = O(n^2(r_{up}^2 + 1)^4 \hat{x^*}^8) = O(n^2(r_{up}^2+1)^4\|x^*\|^8)$ is a bounded random variable. Then, by using Hoeffding's concentration inequality (see (Vershynin, 2018)), we can say that

$$\Pr\left[\left| \frac{1}{m_2}\sum_i \left(\widetilde{X}_i^\top v\right)^4 - \mathbb{E}\left[\left(\widetilde{X}^\top v\right)^4\right]\right| \ge t\right] \le 2\exp\frac{-2m_2t^2}{Cn^4(r_{up}^2+1)^8\|x^*\|^{16}}.$$

Now $\mathbb{E}\left[\left(\widetilde{X}^\top v\right)^4\right] \le \mathbb{E}\left[8(\left(a^\top x^*\right)^8 + \zeta^4)\left(a^\top v\right)^4\right] \overset{(a)}{\le} 131040(K_4^4 + \|x^*\|^8) \le 131040(r_{up}^2+1)^2\|x^*\|^8$, where (a) follows from Equation (13). Now, if we choose $t = 131040(r_{up}^2+1)^2\|x^*\|^8$, by a standard epsilon-net argument, we can say that

$$\Pr\left[\max_{v:\|v\|=1} \sqrt{\frac{1}{m_2}\sum_i \left(\widetilde{X}_i^\top v\right)^4} \ge 724(r_{up}^2+1)\|x^*\|^4\right] \le 2\cdot 12^n \exp\frac{-2m_2}{Cn^4(r_{up}^2+1)^4},$$

where 12^n is the size of the net over n dimensions unit sphere. If $m_2 = \Omega(n^5(r_{up}^2+1)^4\log(4/\delta))$ then, with probability at least $1-\delta/4$, we can say that

$$\max_{w \in \mathfrak{S}_\epsilon^{m_2}} \left\| \sum_i w_i \widetilde{X}_i \widetilde{X}_i^\top - \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i \widetilde{X}_i^\top \right\|_{\text{op}} \le \sqrt{\epsilon}\sqrt{2}\cdot 724(r_{up}^2+1)\|x^*\|^4.$$

Lemma 4.4 implies $14\|x^*\|^4 + \sigma^2 \le \|\widetilde{\Sigma}_r\|_{\text{op}} \le 16\|x^*\|^4 + \sigma^2$. Let's define $\alpha := (r_{up}^2+1)\|x^*\|^4/(14\|x^*\|^4+\sigma^2)$. Then, $1/14 \le \alpha \le (r_{up}^2+1)/14$. So, using this, we can say that with probability at least $1-\delta/4$, the following holds:

$$\max_{w \in \mathfrak{S}_\epsilon^{m_2}} \left\| \sum_i w_i \widetilde{X}_i \widetilde{X}_i^\top - \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i \widetilde{X}_i^\top \right\|_{\text{op}} \le \sqrt{\epsilon}\sqrt{2}\cdot 724\alpha\|\widetilde{\Sigma}_r\|_{\text{op}} \le 74\sqrt{\epsilon}(r_{up}^2+1)\|\widetilde{\Sigma}_r\|_{\text{op}}.$$

Now, for the above choice of ϵ, $74\sqrt{\epsilon}(r_{up}^2+1) \leq 5/(2\cdot(27^2(16+r_{up}^2))) < 1$. Now, by using Theorem F.4, we can conclude that with probability at least $1-\delta/4$ the following holds:

$$\left\| \frac{1}{m_2}\sum_{i=1}^{m_2} \widetilde{X}_i\widetilde{X}_i^\top - \widetilde{\Sigma}_r \right\|_{\text{op}} \leq \frac{5}{2\cdot(27^2(16+r_{up}^2))}\|\widetilde{\Sigma}_r\|_{\text{op}},$$

given that $m_2 \geq \Omega(n\log(4n/\delta)((16+r_{up}^2)^4)$. So, this implies that with probability at least $1-\delta/2$, we can conclude that

$$\max_{w\in\mathfrak{S}_\epsilon^{m_2}} \left\| \sum_i w_i\widetilde{X}_i\widetilde{X}_i^\top - \widetilde{\Sigma}_r \right\|_{\text{op}} \leq \frac{5}{27^2(16+r_{up}^2)}\|\widetilde{\Sigma}_r\|_{\text{op}}.$$

□

F.4 Theorem F.5 and its proof

Theorem F.5. *(Spectral initialization with stability-based Robust PCA) Let $\delta\in[0,0.5]$. Let's assume that an upper bound on signal to noise ratio is known $K_4/\|x^*\|^2 \leq r_{up}$. Under the setting and notation of Algorithm A_{alt}, if $m_1 \geq C_1\log(2/\delta)(1+r_{up}^2)$, $m_2 \geq C_2 n^5(16+r_{up}^2)^4\log(4n/\delta)$, and $\epsilon \leq C_3(16+r_{up}^2)^{-4}$ where C_1, C_2 and C_3 are universal constants. Then with probability at least $0.99-\delta$, the output x_0 of the algorithm A_{alt} satisfies the following:*

$$dist\,(x_0,x^*) \leq \|x^*\|/9. \tag{19}$$

The time complexity of the algorithm is $\widetilde{O}(T_{\text{rob-mean}}(m_1,1)+m_2n)$.

Before presenting the proof of Theorem F.5, we first establish a helper theorem that analyzes the second step of the A_{alt} algorithm. In particular, this helper theorem focuses on the eigenvector direction computation step of A_{alt}. This step is the only difference from Algorithm 1; all the remaining steps are identical, as discussed earlier.

Theorem F.6. ***(Estimating** $\hat{u}$**).** *Assume that $|\hat{\|x^*\|}-\|x^*\|| \leq \frac{1}{27}\|x^*\|$. Let $\delta\in[0,1]$. Under the assumptions and notations from above if $m_2=\max\{n\log(4n/\delta)((16+r_{up}^2)^4), n^5(r_{up}^2+1)^4\log(4/\delta)\}$, and $\epsilon \leq C_2(16+r_{up}^2)^{-4}$ where C_1, C_2 are universal constants. Then, we can say that at least with probability $0.99-\delta/2$,*

$$\text{dist}(\hat{u},u) \leq \frac{1}{27},$$

where u is the top normalized eigenvector of $\widetilde{\Sigma}_r$ matrix.

Proof. Under the assumption of sample size and corruption level, Lemma F.3 implies that the uniform distribution over the uncorrupted truncated samples $\{\widetilde{X}_1,\widetilde{X}_1\ldots,\widetilde{X}_{m_2}\}$ satisfies the stability condition (Definition 4.6) with probability at least $1-\delta/2$ with the $\gamma=5/(27^2(16+r_{up}^2))$. We know that $\|\widetilde{X}\| \leq \hat{\tau}$. Lemma 4.4 implies

$$\|\Sigma-\widetilde{\Sigma}_r\|_{\text{op}} = O(n(K_4^2+\|x^*\|^4)^2/\hat{\tau}^2) \leq \|x^*\|^4,$$

and

$$-\|x^*\|^4\cdot I \preceq (\widetilde{\Sigma}_r)-\Sigma \preceq \|x^*\|^4\cdot I.$$

Now, we know that the top eigenvalue of Σ is $15\|x^*\|^2+\sigma^2$, and the rest of the eigenvalues are the same, with a value of $3\|x^*\|^4+\sigma^2$. Assume that $\widetilde{\Sigma}_r$ has the eigenvalue decomposition $\widetilde{\Sigma}_r = \sum_{i=1}^n \lambda_i f_i f_i^\top$, with $\lambda_1=\lambda_{\max}(\widetilde{\Sigma}_r), f_1=u$ such that the eigenvalues of $\widetilde{\Sigma}_r$ are arranged in non-increasing order, where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A. The Min-max theorem implies that $\lambda_1\in[14\|x^*\|^4+\sigma^2, 16\|x^*\|^4+\sigma^2)]$ and $\lambda_2\in[2\|x^*\|^4+\sigma^2, 4\|x^*\|^4+\sigma^2]$. Now, $4\|x^*\|^4+\sigma^2 \neq 14\|x^*\|^4+\sigma^2$ implies that $\lambda_1\neq\lambda_2$.

$$\begin{aligned}\|\widetilde{\Sigma}_r\|_{\text{op}}-(\hat{u}^\top(\widetilde{\Sigma}_r)\hat{u}) &\geq \lambda_1-(\lambda_1\langle\hat{u},u\rangle^2+(1-\langle\hat{u},u\rangle^2)\lambda_2) = (\lambda_1-\lambda_2)(1-\langle\hat{u},u\rangle^2)\\ &= 10\|x^*\|^4(1-\langle\hat{u},u\rangle^2).\end{aligned}$$

So, now by using theorem Theorem F.1, we can conclude that with probability at least $0.99 - \delta/2$, the following holds:

$$\|\widetilde{\Sigma}_r\|_{\mathrm{op}} - (\hat{u}^\top(\widetilde{\Sigma}_r)\hat{u}) \le O(\gamma)\|\widetilde{\Sigma}_r\|_{\mathrm{op}},$$

which further implies

$$10\|x^*\|^4(1 - \langle \hat{u}, u\rangle^2) \le O(\gamma)(16\|x^*\|^4 + \sigma^2) \le O(\gamma)(16 + r_{up}^2)\|x^*\|^4 \le 5\|x^*\|^4/27^2$$

, which again concludes that

$$|\langle \hat{u}, u\rangle| \ge \sqrt{1 - \frac{1}{2\times 27^2}} \ge \left(1 - \frac{1}{2\times 27^2}\right).$$

Now,

$$\begin{aligned}\mathrm{dist}(\hat{u}, u) &= \min\{\|\hat{u} - u\|_2, \|\hat{u} + u\|_2\} \\ &= \sqrt{2 - 2|\langle \hat{u}, u\rangle|} = \sqrt{2}\sqrt{1 - |\langle \hat{u}, u\rangle|} \le \sqrt{2}\sqrt{1 - \left(1 - \frac{1}{2\times 27^2}\right)} = \frac{1}{27}.\end{aligned}$$

So, from above we can conclude that $\mathrm{dist}(\hat{u}, u) \le 1/27$. So, this implies that $\|\hat{u} - u\| = O(1/27) \le \frac{C}{27}$. If the constant C hidden in the big-$O(\cdot)$ notation is less than 1, then we are done. Otherwise, we set $\gamma = \frac{5}{C^2\, 27^2\, (r_{\mathrm{up}}^2 + 16)}$, and proceed to prove Lemma F.3 using this choice of γ.

□

Proof of the Theorem F.5

Proof. Now, the proof of the Theorem F.5 is analogous to Theorem 4.5 once we have shown Theorem F.6. So, we are omitting the proof here. Note that the time complexity of algorithm A_{alt} follows from the Theorem F.1 and Theorem 3.1. □

G ROBUST PHASE RETRIEVAL UNDER NON-ZERO-MEAN NOISE

As discussed in Section 5, after applying idea of *symmetrization-based pre-processing* from Buna and Rebeschini (2025), the new model is of the form:

$$\upsilon = \langle b, x^*\rangle \cdot \langle c, x^*\rangle + \zeta,$$

where: $b, c \sim \mathcal{N}(0, I_d)$ are independent, $\mathbb{E}[\zeta \mid b, c] = 0, \mathbb{E}[\zeta^2 \mid b, c] = \sigma^2/2, \mathbb{E}[\zeta^4 \mid b, c] \le 2K_4^4$.

We construct a new sample $S' = \{(b_j, c_j, \upsilon_j)\}_{j=1}^m$ from the original sample $S = \{(a_j^*, y_j)\}_{j=1}^{2m}$ by defining: $b_j := a_j^* + a_{m+j}^*/\sqrt{2}, \quad c_j := a_j^* - a_{m+j}^*/\sqrt{2}, \quad \upsilon_j := y_j - y_{m+j}/2.$

Importantly, if the original dataset is ε-corrupted, the new dataset becomes 2ε-corrupted. Now, we will state some statistical properties of the above-defined random variables, which will be needed later.

Proposition G.1.

$$\mathbb{E}[\upsilon] = 0, \mathbb{E}[\upsilon b] = 0, \mathbb{E}[\upsilon^2] = \|x^*\|^4 + \sigma^2/2, \tag{20}$$

$$\mathrm{Var}(\upsilon^2) \le 72(\|x^*\|^8 + K_4^4), \mathbb{E}[\upsilon b^\top c] = \|x^*\|^2, \tag{21}$$

$$\mathrm{Var}(\upsilon b^\top c) = n\|x^*\|^4 + n(\sigma^2/2) + 7\|x^*\|^4, \tag{22}$$

$$\mathrm{Cov}[\upsilon b] = (\|x^*\|^4 + \sigma^2/2)\mathbb{I}_n + 2\|x^*\|^2 x^* x^{*\top}. \tag{23}$$

As discussed in the paragraph titled "Spectral initialization and gradient descent" in the Techniques section, the success of the two-step procedure, namely spectral initialization followed by gradient descent, relies on the local geometry of the population risk. In particular, the population risk

$$r(x) := \tfrac{1}{2}\mathbb{E}[(\upsilon - \langle b, x\rangle\langle c, x\rangle)^2]$$

is strongly convex and smooth within a small neighborhood around x^*. This property (see Proposition G.2 and Proposition G.3 for details) ensures that, once initialized sufficiently close to x^*, gradient descent converges to the true parameter. Consequently, combining robust spectral initialization with robust gradient descent leads to successful recovery. We start by verifying the necessary properties of the population risk function.

G.1 Local properties of population risk for non-zero mean noise robust phase retrieval

We first verify the necessary geometric properties of the population risk $r(x)$ in the neighborhood of $\pm x^*$. These results are given in (Buna and Rebeschini, 2025). We are reproving them for the sake of completeness.

Proposition G.2.

$$r(x) = \mathbb{E}[(v - \langle b, x\rangle\langle c, x\rangle)^2]/2 = \left(\|x^*\|^4 + \|x\|^4 - 2\langle x, x^*\rangle^2 + \sigma^2/2)\right)/2 \tag{24}$$

$$\nabla r(x) = 2x\|x\|^2 - 2\langle x, x^*\rangle x^* \tag{25}$$

$$\nabla^2 r(x) = 2\|x\|^2 \mathbb{I}_n + 4xx^\top - 2x^* x^{*\top}. \tag{26}$$

Proof.

$$\begin{aligned}
2r(x) &= \mathbb{E}[(v - \langle b, x\rangle\langle c, x\rangle)^2] = \mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle + \zeta - \langle b, x\rangle\langle c, x\rangle)^2] \\
&= \mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle - \langle b, x\rangle\langle c, x\rangle)^2] + \mathbb{E}[\zeta^2] + 2\mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle - \langle b, x\rangle\langle c, x\rangle)\zeta] \\
&= \mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle - \langle b, x\rangle\langle c, x\rangle)^2] + \sigma^2/2 + 2\mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle - \langle b, x\rangle\langle c, x\rangle)^2\mathbb{E}[\zeta \mid b, c]] \\
&\overset{(a)}{=} \mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle - \langle b, x\rangle\langle c, x\rangle)^2] + \sigma^2/2 \\
&= \mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle)^2] + \mathbb{E}[(\langle b, x\rangle\langle c, x\rangle)^2] - 2\mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle\langle b, x\rangle\langle c, x\rangle)^2] + \sigma^2/2 \\
&= \mathbb{E}[\langle b, x^*\rangle^2]\mathbb{E}[\langle c, x^*\rangle^2] + \mathbb{E}[\langle b, x\rangle^2]\mathbb{E}[\langle c, x\rangle^2] - 2\mathbb{E}[\langle b, x^*\rangle\langle b, x\rangle]\mathbb{E}[\langle c, x^*\rangle\langle c, x\rangle] + \sigma^2/2 \\
&= \|x^*\|^4 + \|x\|^4 - 2\langle x, x^*\rangle^2 + \sigma^2/2
\end{aligned}$$

where step (a) follows from the fact that $\mathbb{E}[\zeta \mid b, c] = 0$. $\nabla r(x)$ and $\nabla^2 r(x)$ are obtained by further calculation. □

As noted before, for spectral initialization + robust gradient descent to work the loss should be strongly convex and smooth inside a ball of a non-zero radius around x^*. The specific structure of our transformed phase retrieval model ensures that these properties hold in our setting. Our next proposition is about this.

Proposition G.3. *The population loss $r(x)$ is $8\|x^*\|^2/3-$ strongly convex and $6\|x^*\|^2-$ smooth inside a ball of radius $\|x^*\|/9$ around x^*.*

Proof. Consider $h = x - x^*$, such that $\|h\| \leq R\|x^*\|$. Then, replacing x in the expression of the Hessian in Equation (26) with $x^* + h$, the Hessian becomes

$$\nabla^2 r(x) = 4hh^\top + 4h(x^*)^\top + 4x^* h^\top + 2x^*(x^*)^\top + 2\|h\|^2\mathbb{I}_n + 4(h^\top x^*)\mathbb{I}_n + 2\|x^*\|^2\mathbb{I}_n.$$

Using the triangle and Cauchy–Schwarz inequalities along with the identity $\|uv^\top\|_{\mathrm{op}} = \|u\|\|v\|$, we obtain

$$\lambda_{\max}(\nabla^2 r(x)) = \|\nabla^2 r(x)\|_{\mathrm{op}} \leq 6\|h\|^2 + 12\|x^*\|\|h\| + 4\|x^*\|^2 \leq (6R^2 + 12R + 4)\|x^*\|^2.$$

Thus, the smoothness parameter of $r(x)$ is $(6R^2 + 12R + 2)\|x^*\|^2$ for some choice of R. Now, to calculate the strong convexity parameter, we need to put a lower bound on $\lambda_{\min}(\nabla^2 r(x))$.

$$\begin{aligned}
(x^*)^\top \nabla^2 r(x) x^* &= 4(h^\top x^*)^2 + 12(h^\top x^*)\|x^*\|^2 + 2\|x^*\|^2\|h\|^2 + 4\|x^*\|^4 \geq 12(h^\top x^*)\|x^*\|^2 + 4\|x^*\|^4 \\
&\overset{(a)}{\geq} 4\|x^*\|^2\left(\|x^*\|^2 - 3\|h\|\|x^*\|\right)
\end{aligned}$$

where (a) follows from the Cauchy-Schwarz inequality. Thus, $\lambda_{\min}(\nabla^2 r(x)) \geq 4(1 - 3R)\|x^*\|^2$. The conclusion of the theorem follows by choosing $R = \frac{1}{9}$, noting that $6R^2 + 12R + 4 \leq 6$ in this case. □

G.2 Main results for non-mean zero noise

In this section, we first present the formal version of Theorem 5.1 again, and then progressively build up the necessary tools and results required to prove it.

Theorem 5.1. *Consider the same settings as Theorem 4.1. There exists a polynomial-time algorithm (namely, Algorithm 2 (ours work) combined with Algorithm 4 (robust gradient descent) from (Buna and Rebeschini, 2025)) such that, if the total number of samples is* $m = 2m_1 + 2m_2 + 2P\tilde{m}$*, and the following conditions hold:* $m_1 \geq C_2 \log(2/\delta)(1 + r_{up}^2), m_2 \geq C_3 {}^{n(r_{up}^2+1)^2 \log(2n/(\delta\epsilon))}/_{\epsilon}, 2\epsilon \leq {}^{C_1}/_{(r_{up}^2+1)^2}, \quad \tilde{m} \geq C_4 \cdot \max\{n \log n,\ \log(1/\delta)\} \cdot r_{up}^2$*, then, with probability at least* $1-(P+1)\delta$*, the algorithm outputs* x_P *satisfying Equation* (5)*. Moreover, the time complexity of the algorithm is* $O(T_{\text{rob-mean}}(m_1, 1) + m_2^2 n + PT_{\text{rob-mean}}(\tilde{m}, n))$*.*

We will use the following result (Theorem 6.3) of Buna and Rebeschini (2025) who showed that there exists a polynomial time algorithm (Algorithm 4: robust gradient descent by (Buna and Rebeschini, 2025)) with sample complexity $\tilde{O}(n)$ for blind deconvolution problem if given a point x_0 inside a ball of radius $\|x^*\|/9$ centered at $\pm x^*$.

Theorem G.4. ***(Gradient descent for blind deconvolution, see Theorem 6.3 from (Buna and Rebeschini, 2025).)*** *There exists a polynomial time algorithm (see Algorithm 4 from (Buna and Rebeschini, 2025)) that, given a point* $x_0 \in B(\pm x^*, \|x^*\|/9)$ *and if* $2\epsilon \leq \frac{C_1\|x^*\|^4}{\sigma^2}$ *and* P *draws of* $2\tilde{m}$ *samples where* $P\tilde{m} \geq C_2 P \cdot \max\{n \log n, \log(1/\delta)\} \cdot \frac{\sigma^2}{\|x^*\|^4}$*, with probability at least* $1 - P\delta$*, outputs* x_P *satisfying*

$$\frac{dist(x_P, x^*)}{\|x^*\|} \leq C_3 \exp\left(-C_4 P(1-\sqrt{\epsilon})\right) + C_3 \frac{\sigma}{\|x^*\|^2}\left(\sqrt{\frac{n \log n}{\tilde{m}}} + \sqrt{\frac{\log(1/\delta)}{\tilde{m}}}\right) + C_3 \frac{\sigma}{\|x^*\|^2}\sqrt{\epsilon}.$$

In the following Section G.3, we show how to obtain such a point x_0 which gives the desired algorithm together with Theorem G.4.

G.3 Robust spectral initialization for non-mean zero case

Proposition G.1 tells us that x^* is the top eigenvector of $\Sigma := \text{Cov}(vb)$. Motivated by Section 4, define the random vector $X := vb$ and let τ be the truncation parameter, which we will fix later.

Remark G.5. In Algorithm 1, the truncation parameter explicitly depends on an estimate of $\|x^*\|$. The natural idea of estimating $\|x^*\|$ stems from the fact that $\mathbb{E}[vb^\top c] = \|x^*\|^2$ (Equation (21)), which has also appeared in the work of Buna and Rebeschini (2025).[6] Since $\text{Var}(vb^\top c) = \Theta(n)$ (Equation (22)) and the upper bound on the robust estimation error given by Theorem 3.1 is $O(\sqrt{\|\Sigma_{vb^Tc}\|_{\text{op}}\epsilon}) = O(\sqrt{\text{Var}(vb^Tc)\epsilon}) = O(\sqrt{n\epsilon})$, it implies that obtaining a sufficiently accurate approximation of $\|x^*\|^2$ would require a contamination tolerance below $1/n$[7]. Hence, $\|x^*\|$ is not easy to directly estimate accurately, unlike in phase retrieval, and a modification to Algorithm 1 is needed for robust blind deconvolution.

Our approach: We first get a crude estimate of $\|x^*\|^4$ by using the fact that $\mathbb{E}[v^2] = \|x^*\|^4 + \sigma^2/2$ and a known upper bound on the noise to signal ratio, and use it to get the truncation parameter (Lemma G.6). We then apply robust PCA (Theorem 3.2) to the truncated random variable vb to obtain accurate estimates of both the top eigendirection (parallel to x^*) and the top eigenvalue $3\|x^*\|^4 + \sigma^2/2$ of $\Sigma = \text{Cov}(vb)$. Together with an accurate estimate of $\mathbb{E}[v^2] = \|x^*\|^4 + \sigma^2/2$, this yields an accurate estimate of $\|x^*\|^4$. At last, we combine direction and scaling to get the desired output.

Lemma G.6. ***(Truncation.)*** *Define* $\widetilde{X} := X\mathbb{I}_{\|X\|\leq\tau}$*,* $\widetilde{\Sigma}_r := \mathbb{E}[\widetilde{X}\widetilde{X}^\top]$*, where* τ *is the truncation parameter. Then, the following holds:*

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\text{op}} = O(\sqrt{n}(\sigma^2/2 + \|x^*\|^4)^{1/2}(K_4^2 + \|x^*\|^4)/\tau). \tag{27}$$

[6] In Buna and Rebeschini (2025), step 3 of Procedure 3 , they used Theorem 3.1 to robustly estimate $\mathbb{E}[vb^\top c]$. As noted in Remark G.5, the guarantees of Theorem 3.1 is not strong enough to get a sufficiently accurate robust estimate of $\|x^*\|^2$.

[7] Moreover, other natural first order statistics like $\mathbb{E}[vb], \mathbb{E}[vc]$ and other zero order statistics like $\mathbb{E}[v]$ contains no information about $\|x^*\|$.

Proof. Like before, we will first try to bound the tail probability:

$$p = \mathbb{P}(\|X\| \geq \tau) \leq \mathbb{E}[\|X\|^2]/\tau^2 = \mathbb{E}[(\langle b, x^*\rangle\langle c, x^*\rangle + \zeta)^2\|b\|^2]/\tau^2 \leq \mathbb{E}\left[(\langle b, x^*\rangle^2\langle c, x^*\rangle^2 + \zeta^2)\left(\sum_{j=1}^n b_j^2\right)\right]/\tau^2$$
$$\overset{(a)}{\leq} \left(\|x^*\|^2 \operatorname{tr}(\|x^*\|^2\mathbb{I}_n + 2x^*x^{*\top}) + (n\sigma^2/2)\right)/\tau^2 \leq \frac{(n+2)(\|x^*\|^4 + \sigma^2/2)}{\tau^2},$$

where (a) follows from fact that b, c are independent and $\mathbb{E}[\langle b, x^*\rangle^2 bb^\top] = \|x^*\|^2\mathbb{I}_n + 2x^*x^{*\top}$. Now,

$$\Sigma = \mathbb{E}[XX^\top] = \mathbb{E}[XX^\top \mathbb{I}_{\|X\|\leq\tau}] + \mathbb{E}[XX^\top \mathbb{I}_{\|X\|\geq\tau}] = \widetilde{\Sigma}_r + \mathbb{E}[XX^\top \mathbb{I}_{\|X\|\geq\tau}].$$

Upper bounding

$$\begin{aligned}
\|\mathbb{E}[XX^\top \mathbb{I}_{\|X\|\geq\tau}]\|_{\text{op}} &= \sup_{\|v\|=1} \mathbb{E}[\langle v, X\rangle^2 \mathbb{I}_{\|X\|\geq\tau}] \leq \sup_{\|v\|=1} \sqrt{p}\sqrt{\mathbb{E}[\langle v, X\rangle^4]} \\
&\leq \sup_{\|v\|=1} \sqrt{p}\sqrt{\mathbb{E}[8(\zeta^4 + \langle b, x^*\rangle^4\langle c, x^*\rangle^4)\langle b, v\rangle^4]} \leq \sqrt{p}\sqrt{8(6K_4^4 + 2529\|x^*\|^8)} \\
&= O(\sqrt{p}\sqrt{(K_4^4 + \|x^*\|^8)}) = O(\sqrt{p}(K_4^2 + \|x^*\|^4)).
\end{aligned}$$

Now, $\sqrt{p} = O(\sqrt{n(\sigma^2/2 + \|x^*\|^4}/\tau)$. So, this implies

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\text{op}} = O(\sqrt{p}(K_4^2 + \|x^*\|^4)) = O(\sqrt{n(\sigma^2/2 + \|x^*\|^4)}(K_4^2 + \|x^*\|^4)/\tau).$$

□

We now give the details. Assuming a known upper bound on the noise to signal ratio $K_4/\|x^*\|^2 \leq r_{up}$, the choice of $\tau = O(\sqrt{n}(\sigma^2/2 + \|x^*\|^4)^{1/2}(r_{up}^2 + 1))$ (which can be robustly estimated by applying Theorem 3.1 to the random variable v^2, since $\mathbb{E}[v^2] = \sigma^2/2 + \|x^*\|^4$ and this estimate is sufficiently accurate because $\operatorname{Var}(v^2) = O(\|x^*\|^8)$), and using Theorem H.3 and Lemma G.6, we can say that the top eigenvector u of the truncated covariance matrix $\widetilde{\Sigma}_r$ satisfies $dist(\|x^*\|u, x^*) \leq \frac{2\sqrt{2}\|x^*\|\|\Sigma - \widetilde{\Sigma}_r\|_{\text{op}}}{2\|x^*\|^4} \leq \frac{\|x^*\|}{27}$. Thus, u is close to the direction of x^*, and applying robust PCA (Theorem 3.2) on $\widetilde{X}$ yields a robust estimate $\hat{u}$ of the direction of x^*.

We now accurately estimate $\|x^*\|$. Note that $\|\Sigma\|_{\text{op}} = 3\|x^*\|^4 + \sigma^2/2$. By Theorem 3.2, the robust estimate $\lambda_{\hat{u}}$ of the top eigenvalue of $\widetilde{\Sigma}_r$ satisfies $|\lambda_{\hat{u}} - \|\widetilde{\Sigma}_r\|_{\text{op}}| = O(\sqrt{\epsilon}\|x^*\|^4)$. Combining this with Equation (27) and the triangle inequality yields $|\lambda_{\hat{u}} - \|\Sigma\|_{\text{op}}| \leq |\lambda_{\hat{u}} - \|\widetilde{\Sigma}_r\|_{\text{op}}| + |\|\Sigma\|_{\text{op}} - \|\widetilde{\Sigma}_r\|_{\text{op}}| \leq c\|x^*\|^4$, for as small as needed constant c, by the choice of τ and sufficiently small ϵ. This suggests the robust estimator $\widehat{\|x^*\|} = ((\lambda_{\hat{u}} - \hat{v}^2)/2)^{1/4}$ for $\|x^*\|$, where $\hat{v}^2$ estimates $\mathbb{E}[v^2] = \|x^*\|^4 + \sigma^2/2$. See Algorithm 2 for the full procedure. We now present the final theorem for spectral initialization whose proof can be found in Section G.4.

Theorem G.7. ***(Robust spectral initialization for blind deconvolution.)*** *Under the setting and notation of Algorithm 2, if* $2\epsilon \leq C_1(r_{up}^2+1)^{-2}$, $2m_1 \geq C_2\log(3/\delta)(1+r_{up}^2)$ *and* $2m_2 \geq C_3\left(n(r_{up}^2+1)^2\log(3n/(\delta\epsilon))/\epsilon\right)$. *Then, with probability at least* $1-\delta$, $dist\,(x_0, x^*) \leq \|x^*\|/9$. *The algorithm use* $2m_1$ *samples to estimate* $\mathbb{E}[v^2]$ *and* $2m_2$ *samples to estimate* $\hat{u}$ *and* $\lambda_{\hat{u}}$. . *The time complexity of Algorithm 2 is* $O(T_{\text{rob-mean}}(m_1, 1) + m_2^2 n + m_1 n + m_2 n)$.

In the next subsection (Section G.4), we first formally define Algorithm 2 achieving guarantees of Theorem G.7 and then we prove Theorem G.7.

G.4 Formal version of Algorithm 2 and proof of Theorem G.7

The algorithm. Based on the discussions in Section G.3, we have the following informal algorithm (which assumes a known upper bound r_{up} for $K_4/\|x^*\|^2$):

1. Compute an estimate $\hat{v}$ for $(\|x^*\|^4 + \sigma^2/2)^{1/2}$ by using the fact that $\mathbb{E}[v^2] = \|x^*\|^4 + \sigma^2/2$.
2. Use this estimate (together with r_{up}) to compute the truncation threshold $\tau = O(\sqrt{n}\hat{v}(r_{up}^2 + 1)$.
3. Truncate the random variable $X := vb$ using the truncation parameter τ.

4. Apply a robust PCA algorithm (Algorithm 2 from (Kong et al., 2020), Theorem 3.2) on the truncated data to obtain an estimate $\hat{u}$ and $\hat{\lambda}$ of the top (normalized) eigenvector and top eigenvalue of the matrix $\widetilde{\Sigma}_r$.
5. Calculate $(\hat{\lambda} - \hat{v}^2)/2$ as an robust estimate for $\|x^*\|^4$. Now calculate $\|\hat{x^*}\| = \left((\hat{\lambda} - \hat{v}^2)/2\right)^{1/4}$ as an robust estimate for $\|x^*\|$.
6. Scale $\hat{u}$ using the estimated norm $\|\hat{x^*}\|$ to obtain the output x_0 of the robust spectral initialization.

The formal algorithm for robust spectral initialization in blind deconvolution problem is as follows:

Algorithm 2 Spectral Initialisation.

Inputs: failure probability $\delta \in (0, 1)$, corruption fraction $\epsilon > 0$, an upper bound r_{up} on $K_4/\|x^*\|^2$, access to 4 batches of samples out of which 2 batches are of sizes $m_1 = \Omega(\log(2/\delta)(1 + r_{up}^2))$ and remaining batches are of sizes $m_2 = \Omega\left(n(r_{up}^2 + 1)\log(2n/(\delta\epsilon))/\epsilon\right)$ respectively from Model 1.2.
Output: $x_0 \in \mathbb{R}^n$

1: We take two batches of data of size m_1, each consisting of samples of the form (a_i, y_i), and apply the symmetrization trick to obtain a new dataset of size m_1 of the form (v_i, b_i, c_i). Analogously, we perform the same procedure on the other two batches of data of size m_2.
2: $(\|x^*\|^4 + \sigma^2/2)^{1/2}$ **calculation**:
- Receive a set of samples $T_y = \{(b_j)\}_{j=1}^{m_1}$ of size m_1 from the step 1. Form the dataset $T_y^2 = \left\{\left(b_j^2\right)\right\}_{j=1}^{m_1}$.
- Let $\hat{v}^2$ be the robust mean estimate of the dataset T_y^2 by Theorem 3.1.

3: **Robust PCA and robust eigenvalue**:
- Calculate the truncation parameter $\hat{\tau} = O(\sqrt{n}(r_{up}^2 + 1)\hat{v})$.
- Receive a set of samples $T = \{(v_j, b_j, c_j)\}_{j=1}^{m_2}$ of size m_2 from step 1.
- Form the truncated dataset $\widetilde{D} = \{\widetilde{X}_j = X_j \mathbb{I}_{\|X_j\| \le \hat{\tau}} : X_j = v_j b_j\}_{j=1}^{m_2}$.
- Let $\hat{u}$ and $\hat{\lambda}$ be the output of the robust PCA algorithm (Algorithm 2 from (Kong et al., 2020)) applied to the truncated dataset $\widetilde{D}$.

4: $\|x^*\|$ **calculation**:
- Calculate $\|\hat{x^*}\| = \left((\hat{\lambda} - \hat{v}^2)/2\right)^{1/4}$.

5: **Scaling**:
- Return $x_0 = \|\hat{x^*}\|\hat{u}$.

We analyze all the steps of Algorithm 2 sequentially and, finally, combine the results of Steps 2, 3, and 4 of Algorithm 2 to provide the proof of Theorem G.7. We first analyze the step 2 of Algorithm 2.

Theorem G.8. ***(Step 2 of Algorithm 2: Estimate** $(\|x^*\|^4 + \sigma^2/2)^{1/2}$ **using a robust mean estimator).*** *If* $m_1 = \Omega\left(\log(4/\delta)\left(1 + r_{up}^2\right)^2\right)$ *and* $\epsilon = O\left(\left(1 + r_{up}^2\right)^{-2}\right)$, *then with probability at least* $1 - \delta/4$ *we can conclude that*

$$|\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)| = O(\|x^*\|^4) = \|x^*\|^4/54 \tag{28}$$

$$|\hat{v} - (\|x^*\|^4 + \sigma^2/2)^{1/2}| = O(\|x^*\|^2) = \|x^*\|^2/54. \tag{29}$$

Proof. Note that $\mathbb{E}[v^2] = \|x^*\|^4 + \sigma^2/2$ and $\mathrm{Var}(v^2) \le 72(r_{up}^4 + 1)\|x^*\|^8$. Then, using Theorem 3.1 we can say that, with probability at least $1 - \delta/4$:

$$\left|\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)\right| = O\left(\sqrt{\mathrm{Var}(v^2)}\left(\sqrt{2\epsilon} + \sqrt{\frac{\log(4/\delta)}{m_1}}\right)\right) = O\left(\sqrt{(r_{up}^4 + 1)\|x^*\|^8}\left(\sqrt{2\epsilon} + \sqrt{\frac{\log(4/\delta)}{m_1}}\right)\right)$$
$$= O\left((r_{up}^2 + 1)\|x^*\|^4\left(\sqrt{2\epsilon} + \sqrt{\frac{\log(4/\delta)}{m_1}}\right)\right).$$

Now, this implies

$$|\hat{v} - (\|x^*\|^4 + \sigma^2/2)^{1/2}| = \frac{\left|\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)\right|}{(\hat{v} + (\|x^*\|^4 + \sigma^2/2)^{1/2})} \le \frac{\left|\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)\right|}{(\|x^*\|^4 + \sigma^2/2)^{1/2}} \le \frac{\left|\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)\right|}{\|x^*\|^2}$$
$$= O\left((r_{up}^2 + 1)\,\|x^*\|^2 \left(\sqrt{2\epsilon} + \sqrt{\frac{\log(4/\delta)}{m_1}}\right)\right).$$

□

Now we proceed to the analysis of Step 3 of Algorithm 2.

Theorem G.9. ***(Step 3 of Algorithm 1: Estimating** $\hat{u}$**.)*** *Let's assume that* $|\hat{v} - (\|x^*\|^4 + \sigma^2/2)^{1/2}| \le \|x^*\|^2/54$*,* $|\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)| \le \|x^*\|^4/54$ *and* $\delta \in (0, 0.5)$ *is a positive small constant. If* $m_2 = \Omega\left((n + n(r_{up}^2+1)^2\sqrt{\epsilon})\log(4n/(\delta\epsilon))/\epsilon\right)$ *and* $\epsilon \in (0, 1/36]$*. Then, under the above notation, with probability at least* $1 - \delta/4$*, the following holds:*

$$\mathrm{dist}(\hat{u}, u) \le \sqrt{2 - 2\sqrt{1 - O\left((r_{up}^2+1)\sqrt{\epsilon}\right)}}, \tag{30}$$

$$|\lambda_{\hat{u}} - \|\widetilde{\Sigma}_r\|_{\mathrm{op}}| = O(\epsilon(r_{up}^2+1) + \sqrt{\epsilon}(r_{up}^2+1))\|x^*\|^4, \tag{31}$$

where u *is the top normalized eigenvector of* $\widetilde{\Sigma}_r$ *and* $\lambda_{\hat{u}}$ *denote the eigenvalue corresponding to the top eigenvector* $\hat{u}$ *of the estimated second-moment matrix. .*

Proof. As before, we invoke Theorem 3.2 to bound the distance between $\hat{u}$ and the top eigenvector u of $\widetilde{\Sigma}_r$. In order to apply Theorem 3.2, it is necessary to verify that all the stated assumptions are satisfied. We now proceed to check these assumptions one by one. We know that $\|\widetilde{X}\| \le \hat{\tau}$. Note that $|\hat{v} - (\|x^*\|^4 + \sigma^2/2)^{1/2}| \le \|x^*\|^2/54$ implies that $|\hat{v} - (\|x^*\|^4 + \sigma^2/2)^{1/2}| \le (\|x^*\|^4 + \sigma^2/2)^{1/2}/54$, which further implies that $\frac{\|x^*\|^4 + \sigma^2/2)^{1/2}}{\hat{v}} \le 54/53$. Lemma G.6 implies

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\mathrm{op}} = O(\sqrt{n}(\|x^*\|^4 + \sigma^2/2)^{1/2}(r_{up}^2+1)\|x^*\|^4/\hat{\tau}) \le \|x^*\|^4/54 \quad \text{and} \quad -\frac{\|x^*\|^4}{54}I \preceq \widetilde{\Sigma}_r - \Sigma \preceq \frac{\|x^*\|^4}{54}I.$$

The first assumption is to show that $\|\widetilde{X}\widetilde{X}^\top - (\widetilde{\Sigma}_r)\|_{\mathrm{op}} \le B$ for all $\widetilde{X}$ with probability one.

$$\|\widetilde{X}\widetilde{X}^\top - (\widetilde{\Sigma}_r)\|_{\mathrm{op}} \le \|\widetilde{X}\|^2 + \lambda_{max}(\widetilde{\Sigma}_r) \le \hat{\tau}^2 + \lambda_{max}(\Sigma) + \|x^*\|^4 = O(n(r_{up}^2+1)^2\hat{v}^2 + \left(4\|x^*\|^4 + \sigma^2\right))$$
$$= O\left(n(r_{up}^2+1)^2(\|x^*\|^4 + \sigma^2)\right) + (4\|x^*\|^4 + \sigma^2) = O\left(n(r_{up}^2+1)^2(\|x^*\|^4 + \sigma^2)\right).$$

The second assumption is to show that $\mathbb{E}\left[\langle xx^\top, \widetilde{X}\widetilde{X}^\top - (\widetilde{\Sigma}_r)\rangle^2\right] \le v'^2$.

Consider $\|x\| \le 1$, then

$$\mathbb{E}\left[\langle xx^\top, \widetilde{X}\widetilde{X}^\top - (\widetilde{\Sigma}_r)\rangle^2\right] = \mathbb{E}\left[(\langle x, \widetilde{X}\rangle^2 - x^\top\widetilde{\Sigma}_r x)^2\right] = \mathbb{E}\left[(\langle x, \widetilde{X}\rangle)^4\right] - (x^\top\widetilde{\Sigma}_r x)^4 \overset{(a)}{\le} \mathbb{E}\left[(\langle x, X\rangle)^4\right]$$
$$\le \mathbb{E}[8(\zeta^4 + \langle b, x^*\rangle^4\langle c, x^*\rangle^4)\langle b, x\rangle^4] \overset{(b)}{=} O(K_4^4 + \|x^*\|^8) = O((r_{up}^2+1)^2\|x^*\|^8),$$

where (a) follows by applying the trivial upper bound $\hat{\tau} \le \infty$ and (b) follows from Lemma H.2.

This implies

$$\max_{x:\|x\|\le 1} \mathbb{E}\left[\langle xx^\top, \widetilde{X}\widetilde{X}^\top - \widetilde{\Sigma}\rangle^2\right] = O((r_{up}^2+1)^2\|x^*\|^8).$$

In the notations of Theorem 3.2, we have

$$B = O\left(n\,(r_{\mathrm{up}}^2+1)^2(\|x^*\|^4 + \sigma^2)\right), \quad v' = O(\,(r_{\mathrm{up}}^2+1)\,\|x^*\|^4).$$

To apply Theorem 3.2, we must ensure that its sample complexity condition is satisfied. This requires

$$m_2 = \Omega\left(\frac{n+\frac{B\sqrt{\epsilon}}{v'}}{\epsilon}\cdot\log\left(\frac{4n}{\delta\epsilon}\right)\right) = \Omega\left(\frac{n+n\,(r_{\text{up}}^2+1)^2\sqrt{\epsilon}}{\epsilon}\cdot\log\left(\frac{4n}{\delta\epsilon}\right)\right).$$

So, we have verified all assumptions of Theorem 3.2. So, by using Theorem 3.2 and Theorem D.2, we can conclude that, with probability at least $1-\delta/4$, the output $\hat{u}\in\mathbb{R}^{n\times 1}$ and $\lambda_{\hat{u}}$ of Algorithm 2 satisfies:

$$\mathrm{Tr}[P_1(\widetilde{\Sigma}_r)] - \mathrm{Tr}[\hat{u}^\top(\widetilde{\Sigma}_r)\hat{u}] = O\left(\epsilon\cdot\mathrm{Tr}[P_1(\widetilde{\Sigma}_r)] + \sqrt{\epsilon}\cdot v'\right), \quad |\lambda_{\hat{u}} - \|\widetilde{\Sigma}_r\|_{\text{op}}| \le 106\epsilon\|\widetilde{\Sigma}_r\|_{\text{op}} + 212\sqrt{\epsilon}v'.$$

We know that the top eigenvalue of Σ is $3\|x^*\|^4+\sigma^2/2$, while all remaining eigenvalues equal $\|x^*\|^4+\sigma^2/2$. Let $\widetilde{\Sigma}_r = \sum_{i=1}^n \lambda_i f_i f_i^\top$ be the eigenvalue decomposition of $\widetilde{\Sigma}_r$, with $\lambda_1 = \lambda_{\max}(\widetilde{\Sigma}_r)$ and $f_1 = u$ such that eigenvalues of $\widetilde{\Sigma}_r$ are arranged in non-increasing order, where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A. By the min–max theorem, we have $\lambda_1 \in [(161/54)\|x^*\|^4+\sigma^2/2,\ (163/54)\|x^*\|^4+\sigma^2/2]$ and $\lambda_2\in[(53/54)\|x^*\|^4+\sigma^2/2,\ (55/54)\|x^*\|^4+\sigma^2/2]$. Since $(55/54)\|x^*\|^4+\sigma^2/2 \neq (161/4)\|x^*\|^4+\sigma^2/2$, it follows that $\lambda_1\neq\lambda_2$.

$$\begin{aligned}\mathrm{Tr}[P_1(\widetilde{\Sigma}_r)] - \mathrm{Tr}[\hat{u}^\top(\widetilde{\Sigma}_r)\hat{u}] &\ge \lambda_1 - (\lambda_1\langle\hat{u},u\rangle^2 + (1-\langle\hat{u},u\rangle^2)\lambda_2) = (\lambda_1-\lambda_2)(1-\langle\hat{u},u\rangle^2)\\ &= 106/54(1-\langle\hat{u},u\rangle^2).\end{aligned}$$

So now,

$$(1-\langle\hat{u},u\rangle^2) = O\left(\epsilon\left(1+\frac{\sigma^2}{\|x^*\|^4}\right) + 73(r_{up}^2+1)\sqrt{\epsilon}\right) = O\left(\epsilon(r_{up}^2+1) + (r_{up}^2+1)\sqrt{\epsilon}\right) = O\left((r_{up}^2+1)\sqrt{\epsilon}\right).$$

This implies

$$|\langle\hat{u},u\rangle| \ge \sqrt{1-O\left((r_{up}^2+1)\sqrt{\epsilon}\right)}.$$

Now, we can conclude that,

$$\begin{aligned}\mathrm{dist}(\hat{u},u) &= \min\{\|\hat{u}-u\|_2, \|\hat{u}+u\|_2\}\\ &= \sqrt{2-2|\langle\hat{u},u\rangle|}\\ &\le \sqrt{2-2\sqrt{1-O\left((r_{up}^2+1)\sqrt{\epsilon}\right)}}.\end{aligned}$$

and

$$|\lambda_{\hat{u}} - \|\widetilde{\Sigma}_r\|_{\text{op}}| \le 106\epsilon\|\widetilde{\Sigma}_r\|_{\text{op}} + 212\sqrt{\epsilon}v' = O(\epsilon(r_{up}^2+1)+\sqrt{\epsilon}(r_{up}^2+1))\|x^*\|^4.$$

□

Now we proceed to the analysis of Step 4 of Algorithm 2.

Theorem G.10. ***(Step 4 of Algorithm 2: Estimate** $\|x^*\|$**).*** *If* $m_1 = \Omega\left(\log(4/\delta)\left(1+r_{up}^2\right)^2\right)$, $m_2 = \Omega\left((n+n(r_{up}^2+1)^2\sqrt{\epsilon})\log(4n/(\delta\epsilon))/\epsilon\right)$ *and* $\epsilon = O\left(\left(1+r_{up}^2\right)^{-2}\right)$, *then with probability at least* $1-\delta/2$ *we can conclude that*

$$|\hat{\|x^*\|} - \|x^*\|| \le \frac{\|x^*\|}{27} \tag{32}$$

Proof. Using Theorem G.8, Lemma G.6 and Theorem G.9, we can conclude that with probability at least $1-\delta/2$ the following holds:

$$|\hat{v}^2 - (\|x^*\|^4+\sigma^2/2)| \le \|x^*\|^4/54, \quad \|\Sigma-\widetilde{\Sigma}_r\|_{\text{op}} \le \|x^*\|^4/54,$$

$$|\lambda_{\hat{u}} - \|\widetilde{\Sigma}_r\|_{\text{op}}| = O(\epsilon(r_{up}^2+1)+\sqrt{\epsilon}(r_{up}^2+1))\|x^*\|^4 = \|x^*\|^4/27.$$

This implies

$$\begin{aligned}
|(\lambda_{\hat{u}} - \hat{v}^2)/2 - \|x^*\|^4| &= |(\lambda_{\hat{u}} - (3\|x^*\|^4 + \sigma^2/2))/2 - (\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2))/2)| \\
&\le 1/2|(\lambda_{\hat{u}} - (3\|x^*\|^4 + \sigma^2/2))| + 1/2|(\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)))| \\
&\le 1/2|(\lambda_{\hat{u}} - \|\widetilde{\Sigma}_r\|_{\mathrm{op}}| + 1/2|\|\widetilde{\Sigma}_r\|_{\mathrm{op}} - (3\|x^*\|^4 + \sigma^2/2))| + 1/2|(\hat{v}^2 - (\|x^*\|^4 + \sigma^2/2)))| \\
&\le \frac{1}{2}\left(\frac{\|x^*\|^4}{27} + \frac{\|x^*\|^4}{54} + \frac{\|x^*\|^4}{54}\right) = \frac{\|x^*\|^4}{27}.
\end{aligned}$$

Now, we know that for any $a, b, c > 0$ with $b^2 > c$, $\left|a^2 - b^2\right| \le c$ implies $|a - b| \le b - \sqrt{b^2 - c}$. By using the above property, we get to

$$|\sqrt{(\lambda_{\hat{u}} - \hat{v}^2)/2} - \|x^*\|^2\,| \le \|x^*\|^2 - \sqrt{26/27}\|x^*\|^2 \le \|x^*\|^2/27.$$

Now, using the above result once more we conclude that

$$\left|((\lambda_{\hat{u}} - \hat{v}^2)/2)^{1/4} - \|x^*\|\right| \le \|x^*\| - \sqrt{26/27}\|x^*\| \le \|x^*\|/27.$$

Finally, we can conclude that

$$\left|\widehat{\|x^*\|} - \|x^*\|\right| \le \|x^*\|/27.$$

□

Proof of Theorem G.7.

Now we are ready to give the proof of Theorem G.7.

Proof. Under the assumptions on the sample complexity m_1, m_2 and the corruption level ϵ, and using Theorem G.10, Theorem G.9, Theorem G.8, and Lemma G.6, we conclude that, with probability at least $1 - \delta$, the following holds:

$$\mathrm{dist}(\hat{u}, u) \le \sqrt{2 - 2\sqrt{1 - O\left((r_{up}^2 + 1)\sqrt{\epsilon}\right)}} \le (1/27),$$

$$\|\Sigma - \widetilde{\Sigma}_r\|_{\mathrm{op}} = O(\sqrt{n}(\|x^*\|^4 + \sigma^2/2)^{1/2}(r_{up}^2 + 1)\|x^*\|^4/\hat{\tau} \le \frac{\|x^*\|^4}{54} \le \frac{\sqrt{2}\|x^*\|^4}{54},$$

$$|\widehat{\|x^*\|} - \|x^*\|| \le \|x^*\|/27,$$

where u is top normalized eigenvector of $\widetilde{\Sigma}_r$. Now, by applying Theorem H.3 (Davis-Kahan), we can conclude that

$$\mathrm{dist}(\|x^*\|u, x^*) \le \frac{2\sqrt{2}\|x^*\|\|\Sigma - (\widetilde{\Sigma}_r)\|_{\mathrm{op}}}{2\|x^*\|^4} \le \frac{\|x^*\|}{27} \quad \text{and}$$

$$\mathrm{dist}(-\|x^*\|u, x^*) \le \frac{2\sqrt{2}\|x^*\|\|\Sigma - (\widetilde{\Sigma}_r)\|_{\mathrm{op}}}{2\|x^*\|^4} \le \frac{\|x^*\|}{27}.$$

So,

$$\begin{aligned}
\mathrm{dist}\,(x_0, x^*) = \mathrm{dist}\left(\widehat{\|x^*\|}\hat{u}, x^*\right) &\le \mathrm{dist}(\widehat{\|x^*\|}\hat{u}, \|x^*\|\hat{u}) + \mathrm{dist}(\|x^*\|\hat{u}, \|x^*\|u) + dist\,(\|x^*\|u, x^*) \\
&= \|\widehat{\|x^*\|}\hat{u} - \|x^*\|\hat{u}\| + \mathrm{dist}(\|x^*\|\hat{u}, \|x^*\|u) + dist\,(\|x^*\|u, x^*) \\
&= |\widehat{\|x^*\|} - \|x^*\|| + \|x^*\|\,\mathrm{dist}(\hat{u}, u) + dist\,(\|x^*\|u, x^*) \\
&\le \frac{\|x^*\|}{27} + \frac{\|x^*\|}{27} + \frac{\|x^*\|}{27} = \frac{\|x^*\|}{9}.
\end{aligned}$$

□

G.5 Proof of Theorem 5.1

Theorem 5.1 now follows by combining Theorem G.4 and Theorem G.7.

G.6 Proof of Proposition G.1

Proof.

$$\mathbb{E}[\upsilon^2] = \mathbb{E}\left[\left(b^\top x^*\right)^2\left(c^\top x^*\right)^2 + \zeta^2 + 2\zeta\left(b^\top x^*\right)\left(c^\top x^*\right)\right] = \mathbb{E}\left[\left(b^\top x^*\right)^2\left(c^\top x^*\right)^2 + \zeta^2\right] = \|x^*\|^4 + \sigma^2/2$$

$$\begin{aligned}\mathrm{Var}(\upsilon^2) &= \mathbb{E}\left[\left(\left(b^\top x^*\right)\left(c^\top x^*\right) + \zeta\right)^4\right] - (\|x^*\|^4 + \sigma^2/2)^2 \leq \mathbb{E}\left[8\left(b^\top x^*\right)^4\left(c^\top x^*\right)^4\right] + \mathbb{E}\left[8(\zeta)^4\right] - \|x^*\|^8 \\ &\leq 72\,\|x^*\|^8 + 16K_4^4 - \|x^*\|^8 \leq 72(r_{up}^4 + 1)\,\|x^*\|^8\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\upsilon b] &= \mathbb{E}[\langle b, x^*\rangle \cdot \langle c, x^*\rangle b + \zeta b] = \mathbb{E}[\langle b, x^*\rangle \cdot \langle c, x^*\rangle b] = 0, \mathbb{E}[\upsilon c] = \mathbb{E}[\langle b, x^*\rangle \cdot \langle c, x^*\rangle c + \zeta c] \\ &= \mathbb{E}[\langle b, x^*\rangle \cdot \langle c, x^*\rangle c] = 0\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\upsilon b^\top c] &= \mathbb{E}[\langle b, x^*\rangle \cdot \langle c, x^*\rangle b^\top c + \zeta b^\top c] = \mathbb{E}[\langle b, x^*\rangle \cdot \langle c, x^*\rangle b^\top c] = \mathbb{E}[\sum_{i,j,k=1}^{n} b_j x_j^* c_i x_i^* b_k c_k] \\ &= \sum_{i,j,k=1}^{n} \delta_{\{j=k\}} x_j^* \delta_{\{i=k\}} x_i^* = \sum_{k=1}^{n} x_k^{*2} = \|x^*\|^2,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\upsilon b^\top c) &= \mathrm{Tr}(\mathbb{E}[\langle b, x^*\rangle^2 bb^\top]\mathbb{E}[\langle c, x^*\rangle^2 cc^\top]) + (\sigma^2/2)\mathbb{E}[(b^\top c)^2] - \|x^*\|^4 = \mathrm{Tr}((\|x^*\|^2\,\mathbb{I}_n + 2x^*x^{*\top})^2) \\ &+ (\sigma^2/2)n - \|x^*\|^4 = (8+n)\|x^*\|^4 + (\sigma^2/2)n - \|x^*\|^4 = n\|x^*\|^4 + n(\sigma^2/2) + 7\|x^*\|^4.\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}[\upsilon b] &= \mathbb{E}[\upsilon^2 bb^\top] - \mathbb{E}[\upsilon b]\mathbb{E}[\upsilon b]^\top = \mathbb{E}[\upsilon^2 bb^\top] = \mathbb{E}[\langle b, x^*\rangle^2 \cdot \langle c, x^*\rangle^2 bb^\top] + \mathbb{E}[\zeta^2 bb^\top] = \|x^*\|^2\mathbb{E}[\langle b, x^*\rangle^2 bb^\top] \\ &+ \sigma^2/2\mathbb{I}_n = (\|x^*\|^4 + \sigma^2/2)\mathbb{I}_n + 2\|x^*\|^2 x^*x^{*\top}\end{aligned}$$

. ∎

H NECESSARY LEMMAS

Lemma H.1. *Let's assume that $x^* \in \mathbb{R}^n$ is the fixed vector and $a \sim \mathcal{N}(0, \mathbb{I}_n)$. Then,*

$$\mathbb{E}(\langle a, x^*\rangle^2 aa^T) = 2x^*x^{*T} + ||x^*||^2\mathbb{I}_n, \mathbb{E}(\langle a, x^*\rangle^4 aa^T) = ||x^*||^4\left(3\mathbb{I}_n + \frac{12}{||x^*||^2}x^*x^{*T}\right).$$

Proof. Now, let's write $a = \langle a, x^*\rangle \frac{x^*}{||x^*||^2} + v_1$ such that $\langle v_1, x^*\rangle = 0$. Now define $\alpha_1 := \frac{\langle a, x^*\rangle}{||x^*||^2}$, so we can write $a = \alpha_1 x^* + v_1$. Note that because of the above dissociation, α_1 and v_1 are now independent random variables. So, $\langle a, x^*\rangle^2 = \alpha_1^2||x^*||^4$, and $aa^\top = \alpha_1^2 x^*x^{*\top} + v_1v_1^\top + \alpha_1 x^{*\top}v_1 + \alpha_1 v_1 x^{*\top}$.

$$\begin{aligned}\mathbb{I}_n &= \mathbb{E}[aa^\top] = \mathbb{E}[\alpha_1^2 x^*x^{*\top}] + \mathbb{E}[v_1v_1^\top] + \mathbb{E}[\alpha_1 x^{*\top}v_1] + \mathbb{E}[\alpha_1 v_1 x^{*\top}] \\ &= \mathbb{E}[\alpha_1^2]x^*x^{*\top} + \mathbb{E}[v_1v_1^\top] + x^*\mathbb{E}[\alpha_1 v_1^\top] + \mathbb{E}[\alpha_1 v_1]x^{*\top} \overset{(a)}{=} \frac{1}{||x^*||^2}x^*x^{*\top} + \mathbb{E}[v_1v_1^\top],\end{aligned}$$

where (a) follows from the fact that $x^*\mathbb{E}[\alpha_1 v_1^\top] = 0, \mathbb{E}[\alpha_1 v_1]x^{*\top} = 0$ because $\mathbb{E}[\alpha_1 v_1^\top] = \mathbb{E}[\alpha_1]\mathbb{E}[v_1^\top] = 0$ and $\mathbb{E}[\alpha_1 v_1] = \mathbb{E}[\alpha_1 v_1^\top]^\top$. Now,

$$\begin{aligned}\mathbb{E}[\langle a, x^*\rangle^2 aa^\top] &= \mathbb{E}[\alpha_1^2||x^*||^4(\alpha_1^2 x^*x^{*\top} + v_1v_1^\top)] = ||x^*||^4\mathbb{E}[\alpha_1^4]x^*x^{*\top} + ||x^*||^4\mathbb{E}[\alpha_1^2 v_1v_1^\top] \\ &= ||x^*||^4\frac{3||x^*||^4}{||x^*||^8}x^*x^{*\top} + \frac{||x^*||^4}{||x^*||^2}\left(\mathbb{I}_n - \frac{1}{||x^*||^2}x^*x^{*\top}\right) = 2x^*x^{*T} + ||x^*||^2\mathbb{I}_n\end{aligned}$$

$$\mathbb{E}[\langle a,x^*\rangle^4 aa^\top] = \mathbb{E}[\alpha_1^4||x^*||^8(\alpha_1^2 x^*x^{*\top} + v_1v_1^\top)] = ||x^*||^8\mathbb{E}[\alpha_1^6]x^*x^{*\top} + ||x^*||^8\mathbb{E}[\alpha_1^4 v_1v_1^\top]$$
$$= ||x^*||^8\frac{15||x^*||^6}{||x^*||^{12}}x^*x^{*\top} + ||x^*||^8\frac{3||x^*||^4}{||x^*||^8}\left(\mathbb{I}_n - \frac{1}{||x^*||^2}x^*x^{*\top}\right) = ||x^*||^4\left(3\mathbb{I}_n + \frac{12}{||x^*||^2}x^*x^{*T}\right).$$

□

Lemma H.2. *Let's assume that $v, x^* \in \mathbb{R}^n$ are the fixed vectors and $a \sim \mathcal{N}(0,\mathbb{I}_n)$. If i is even, then*

$$\mathbb{E}\left[\left(a^\top x^*\right)^i\left(v^\top a\right)^4\right] \le (i+3)!!||x^*||^{i-4}(v^\top x^*)^4 + ||x^*||^i(3(i-1)!! + 6(i+1)!!)||v||^4.$$

If i is odd, then

$$\mathbb{E}\left[\left(a^\top x^*\right)^i\left(v^\top a\right)^4\right] = 0.$$

Proof. In the notations of the proof of Lemma H.1, if i is even, then

$$\mathbb{E}\left[\left(a^\top x^*\right)^i\left(v^\top a\right)^4\right] = ||x^*||^{2i}\mathbb{E}\left[\alpha_1^i\left(v^\top a\right)^4\right] = ||x^*||^{2i}\mathbb{E}\left[\alpha_1^i\left(v^\top\alpha_1 x^* + v^\top v_1\right)^4\right] = ||x^*||^{2i}\mathbb{E}[\alpha_1^i(\alpha_1^4(v^\top x^*)^4$$
$$+ (v^\top v_1)^4 + 6\alpha_1^2(v^\top v_1)^2(v^\top x^*)^2 + 4\alpha_1^3(v^\top x^*)^3(v^\top v_1) + 4(v^\top v_1)^3\alpha_1(v^\top x^*))]$$
$$= ||x^*||^{2i}\mathbb{E}[\alpha_1^{(i+4)}(v^\top x^*)^4 + \alpha_1^{(i)}(v^\top v_1)^4 + 6\alpha_1^{i+2}(v^\top v_1)^2(v^\top x^*)^2 + 4\alpha_1^{i+3}(v^\top x^*)^3(v^\top v_1)$$
$$+ 4(v^\top v_1)^3\alpha_1^{i+1}(v^\top x^*))] = ||x^*||^{2i}\mathbb{E}[\alpha_1^{(i+4)}(v^\top x^*)^4 + \alpha_1^{(i)}(v^\top v_1)^4 + 6\alpha_1^{i+2}(v^\top v_1)^2(v^\top x^*)^2]$$
$$= ||x^*||^{2i}\left(\mathbb{E}[\alpha_1^{(i+4)}](v^\top x^*)^4 + \mathbb{E}[\alpha_1^{(i)}]\mathbb{E}[(v^\top v_1)^4] + 6\mathbb{E}[\alpha_1^{(i+2)}]\mathbb{E}[(v^\top v_1)^2](v^\top x^*)^2\right)$$
$$= ||x^*||^{i-4}(i+3)!!(v^\top x^*)^4 + ||x^*||^i(i-1)!!\mathbb{E}[(v^\top v_1)^4] + 6||x^*||^{i-2}(i+1)!!\mathbb{E}[(v^\top v_1)^2](v^\top x^*)^2.$$

Since $\Sigma_{v_1} = \mathbb{I}_n - \frac{1}{||x^*||^2}x^*x^{*\top}$. So, $\mathbb{E}[(v^\top v_1)^2] = ||v||^2 - \frac{(v^\top x^*)^2}{||x^*||^2} \le ||v||^2$.

$$3||v||^4 = \mathbb{E}[(v^\top a)^4] = \mathbb{E}[\alpha_1^4(v^\top x^*)^4 + (v^\top v_1)^4 + 6\alpha_1^2(v^\top v_1)^2(v^\top x^*)^2 + 4\alpha_1^3(v^\top x^*)^3(v^\top v_1) + 4(v^\top v_1)^3\alpha_1(v^\top x^*)]$$
$$= \mathbb{E}[\alpha_1^4(v^\top x^*)^4 + (v^\top v_1)^4 + 6\alpha_1^2(v^\top v_1)^2(v^\top x^*)^2] = \frac{3(v^\top x^*)^4}{\|x^*\|^4} + \mathbb{E}[(v^\top v_1)^4]$$
$$+ 6(v^\top x^*)^2\frac{1}{\|x^*\|^2}\left(||v||^2 - \frac{v^\top x^*)^2}{||x^*||^2}\right) \overset{(a)}{\ge} \mathbb{E}[(v^\top v_1)^4],$$

where (a) follows from the fact that $\frac{3(v^\top x^*)^4}{\|x^*\|^4}$ and $6(v^\top x^*)^2\frac{1}{\|x^*\|^2}\left(||v||^2 - \frac{(v^\top x^*)^2}{||x^*||^2}\right)$ are positive constants. So,

$$\mathbb{E}\left[\left(a^\top x^*\right)^i\left(v^\top a\right)^4\right] = |x^*||^{i-4}(i+3)!!(v^\top x^*)^4 + ||x^*||^i(i-1)!!\mathbb{E}[(v^\top v_1)^4] + 6||x^*||^{i-2}(i+1)!!\mathbb{E}[(v^\top v_1)^2](v^\top x^*)^2$$
$$\le |x^*||^{i-4}(i+3)!!(v^\top x^*)^4 + 3||x^*||^i(i-1)!!||v||^4 + 6||x^*||^{i-2}(i+1)!!||v||^2(v^\top x^*)^2$$
$$\le (i+3)!!|x^*||^{i-4}(v^\top x^*)^4 + ||x^*||^i(3(i-1)!! + 6(i+1)!!)||v||^4.$$

In the notation of the proof of Lemma H.1, if i is odd, then

$$\mathbb{E}\left[\left(a^\top x^*\right)^i\left(v^\top a\right)^4\right] = ||x^*||^{2i}\mathbb{E}\left[\alpha_1^i\left(v^\top a\right)^4\right] = ||x^*||^{2i}\mathbb{E}\left[\alpha_1^i\left(v^\top\alpha_1 x^* + v^\top v_1\right)^4\right]$$
$$= ||x^*||^{2i}\mathbb{E}[\alpha_1^{(i+4)}(v^\top x^*)^4 + \alpha_1^{(i)}(v^\top v_1)^4 + 6\alpha_1^{i+2}(v^\top v_1)^2(v^\top x^*)^2 + 4\alpha_1^{i+3}(v^\top x^*)^3(v^\top v_1)$$
$$+ 4(v^\top v_1)^3\alpha_1^{i+1}(v^\top x^*))] = ||x^*||^{2i}\mathbb{E}[4\alpha_1^{i+3}(v^\top x^*)^3(v^\top v_1) + 4(v^\top v_1)^3\alpha_1^{i+1}(v^\top x^*))]$$
$$= 4||x^*||^{2i}\left(\mathbb{E}[\alpha_1^{(i+3)}](v^\top x^*)^3\mathbb{E}[(v^\top v_1)] + \mathbb{E}[\alpha_1^{(i+1)}]\mathbb{E}[(v^\top v_1)^3]\right)$$
$$= 4||x^*||^{2i}\left(\mathbb{E}[\alpha_1^{(i+1)}]\mathbb{E}[(v^\top v_1)^3]\right),$$

and

$$0 = \mathbb{E}[(v^\top a)^3] = \mathbb{E}[\alpha_1^3(v^\top x^*)^3 + (v^\top v_1)^3 + 3\alpha_1^2(v^\top x^*)^2(v^\top v_1) + 3(v^\top v_1)^2\alpha_1(v^\top x^*)]$$
$$= \mathbb{E}[\alpha_1^3(v^\top x^*)^3 + (v^\top v_1)^3] = \mathbb{E}[(v^\top v_1)^3].$$

So, this implies $\mathbb{E}\left[\left(a^\top x^*\right)^i\left(v^\top a\right)^4\right] = 4||x^*||^{2i}\left(\mathbb{E}[\alpha_1^{(i+1)}]\mathbb{E}[(v^\top v_1)^3]\right) = 0$ □

Theorem H.3. *(See Theorem 4.5.5 and the following conclusion from (Vershynin, 2018), Edition 1.) Let S and T be symmetric matrices with the same dimensions. Fix i and assume that the i-th largest eigenvalue of S is well separated from the rest of the spectrum:*

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0$$

Let's assume that $v_i(S)$ and $v_i(T)$ are the unit eigenvectors corresponding to the i-th largest eigenvalues of the matrices S and T, respectively. Then the unit eigenvectors $v_i(S)$ and $v_i(T)$ are close to each other up to a sign, namely

$$\exists \theta \in \{-1, 1\} : \quad \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{2^{3/2}\|S - T\|}{\delta}$$

Proof. Let $u = v_i(S)$ and $v = v_i(T)$ be unit eigenvectors corresponding to the i-th largest eigenvalues of symmetric matrices S and T, respectively. From the Davis-Kahan theorem (Theorem 4.5.5 from (Vershynin, 2018), Edition 1), we have

$$\sin \angle(u, v) \leq \frac{2\|S - T\|}{\delta}.$$

Let $\theta := \mathrm{sign}(\langle u, v\rangle) \in \{-1, 1\}$ be chosen so that the angle between u and θv lies in $[0, \pi/2]$ because $\cos \angle(u, \theta v) = \theta(\langle u, v\rangle) = |(\langle u, v\rangle)|$. Then,

$$\|u - \theta v\|_2^2 = \langle u - v, u - v\rangle = \|u\|^2 + \|\theta v\|^2 - 2\langle u, \theta v\rangle = 2 - 2\theta\langle u, v\rangle = 2(1 - |\langle u, v\rangle|).$$

From the sine-angle bound,

$$\sin \angle(u, v) = \sqrt{1 - \langle u, v\rangle^2} \leq \frac{2\|S - T\|}{\delta},$$

which implies

$$|\langle u, v\rangle| \geq \sqrt{1 - \left(\frac{2\|S - T\|}{\delta}\right)^2}.$$

Therefore,

$$\|u - \theta v\|_2^2 \leq 2\left(1 - \sqrt{1 - \left(\frac{2\|S - T\|}{\delta}\right)^2}\right).$$

Using the inequality $1 - \sqrt{1 - x} \leq x$ for $x \in [0, 1]$, we get

$$\|u - \theta v\|_2^2 \leq 2\left(\frac{2\|S - T\|}{\delta}\right)^2 = \left(\frac{2^{3/2}\|S - T\|}{\delta}\right)^2,$$

and hence,

$$\|u - \theta v\|_2 \leq \frac{2^{3/2}\|S - T\|}{\delta}.$$

□