Proceedings of Machine Learning Research vol 65:1–44, 2017

Computationally Efficient Robust Sparse Estimation in High Dimensions

Sivaraman Balakrishnan SIVA@STAT.CMU.EDU
Department of Statistics, Carnegie Mellon University
Simon S. Du SSDU@CS.CMU.EDU
Machine Learning Department, School of Computer Science, Carnegie Mellon University
Jerry Li JERRYZLI@MIT.EDU
Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology
Aarti Singh AARTISINGH@CMU.EDU
Machine Learning Department, School of Computer Science, Carnegie Mellon University

Abstract

Many conventional statistical procedures are extremely sensitive to seemingly minor deviations from modeling assumptions. This problem is exacerbated in modern high-dimensional settings, where the problem dimension can grow with and possibly exceed the sample size. We consider the problem of *robust estimation* of sparse functionals, and provide a computationally and statistically efficient algorithm in the high-dimensional setting. Our theory identifies a unified set of deterministic conditions under which our algorithm guarantees accurate recovery. By further establishing that these deterministic conditions hold with high-probability for a wide range of statistical models, our theory applies to many problems of considerable interest including sparse mean and covariance estimation; sparse linear regression; and sparse generalized linear models. In certain settings, such as the detection and estimation of sparse principal components in the spiked covariance model, our general theory does not yield optimal sample complexity, and we provide a novel algorithm based on the same intuition which is able to take advantage of further structure of the problem to achieve nearly optimal rates.

Keywords: Robustness, sparsity, linear regression, covariance estimation, sparse principal components analysis (PCA), generalized linear models, logistic regression.

1. Introduction

Complex high-dimensional datasets pose a variety of computational and statistical challenges. In attempts to address these challenges, the past decade has witnessed a significant amount of research on sparsity constraints in statistical models. Sparsity constraints often lead to more interpretable models, that can be estimated efficiently even in the high-dimensional regime where the sample size n can be dwarfed by the model dimension d.

Much of the theoretical literature on sparse estimation has focused on providing guarantees under strong, often impractical, generative assumptions. This in turn motivates the study of the robustness of these statistical estimators, and the design of new robust estimators. Classically, the sensitivity of conventional statistical procedures to apparently small deviations from the assumed statistical model, was noted by Tukey (1975) who observed that estimators like the empirical mean can be sensitive to even a single gross outlier. The formal study, of robust estimation, was initiated by Huber (1964, 1965) who considered estimation procedures under the ϵ*-contamination model*, where samples are obtained from a mixture model of the form:

$$P_\epsilon = (1-\epsilon)P + \epsilon Q, \tag{1}$$

where P is the uncontaminated target distribution, Q is an arbitrary outlier distribution and ϵ is the expected fraction of outliers. Subsequent work in the literature on robust statistics, focused on the design of robust estimators and the study of their statistical properties (see, for instance, the works of Huber (2011); Hampel et al. (2011)). Recent research (Chen et al., 2015, 2016) has focussed on providing a complementary minimax perspective by characterizing both minimax upper and lower bounds on the performance of estimators in a variety of settings. Notably, the minimax estimation rates in these settings typically have two aspects: (1) the dependence on the contamination parameter ϵ, which we refer to as the *contamination dependence*, and (2) the *statistical rate* (typically, a function of the sample size n and the dimensionality d).

The major drawback of many of these classical robust estimators is that they are either heuristic in nature (for instance, methods based on Winsorization (Hastings Jr et al., 1947)) and are generally not optimal in the minimax sense, or are computationally intractable (for instance, methods based on Tukey's depth (Tukey, 1975) or on ℓ_1 tournaments (Yatracos, 1985)).

Considering the low-dimensional setting where $d \ll n$, recent works (Diakonikolas et al., 2016a; Lai et al., 2016; Charikar et al., 2016) provide some of the first computationally tractable, provably robust estimators with near-optimal contamination dependence in a variety of settings. Concretely, the paper of Lai et al. (2016) considers robust mean and covariance estimation for distributions with appropriately controlled moments, while the work of Diakonikolas et al. (2016a), focuses on robust mean and covariance estimation for Gaussians and extends these results to various other models including the mixture of Gaussians. Although the focus of our paper is on the multivariate setting, we note that several recent papers have provided robust estimation guarantees for univariate distributions (Acharya et al., 2017; Chan et al., 2013, 2014; Daskalakis et al., 2012; Diakonikolas et al., 2016b).

A Unified Framework for Robust Recovery We make several contributions to this line of research. In more details, we focus on the *sparse high-dimensional setting* where the dimensionality d is potentially much larger than the sample-size n but the unknown target parameter is k sparse (and $k \ll n$). Building on the work of Diakonikolas et al. (2016a), our first main contribution is to provide a unified framework for the estimation of sparse functionals. We identify a set of core deterministic conditions, under which we can guarantee accurate recovery of a statistical functional in polynomial-time. In contrast to prior work, this framework unifies, for instance, the robust estimation of the mean vector and of the covariance matrix of a high-dimensional distribution. Our second main contribution, establishes that these deterministic conditions hold with high-probability in many statistical models, even in the high-dimensional setting where $n \ll d$, under appropriate sparsity assumptions. As a consequence, we obtain the first robust estimators in a variety of high-dimensional problems of practical interest including sparse mean and covariance estimation; sparse linear regression; and sparse generalized linear models. Finally, from a technical standpoint, as will be discussed at more length in the sequel we introduce a variety of new techniques involving the careful analysis of convex relaxations and delicate truncation arguments that we anticipate will be useful in other related problems.

At a high level, we show that if the target parameter (e.g. mean or covariance) is corrupted, higher order moments must also be corrupted. Moreover, if the target parameter is sparse, then the corruption of the higher order moments can be detected by inspecting a few coordinates, which in turn allows us to improve our estimate of the target parameter. However, this comes at an inherent statistical cost. In typical (non-robust) sparse estimation problems it suffices to obtain enough samples for the empirical estimate to converge to its sparse population counterpart in the appropriate sparsity-respecting norm. Now, we also need some of the empirical higher order moments to converge to their population counterparts, requiring more samples. This statistical price seems unavoidable for computationally tractable procedures—in recent work, Diakonikolas et al. (2016c) showed any statistical query (SQ) algorithm (Kearns, 1998) for robustly learning the sparse mean of a Gaussian must pay this extra statistical cost. Specifically, they show that (up to log factors) our rates are optimal for any SQ algorithm.

Going Beyond the Framework This poses the natural question: in what settings can we avoid this statistical blowup? We consider the problem of estimating high-dimensional principal components in the *spiked covariance model* (Johnstone, 2001) and show that by going beyond our general framework and by tailoring our estimator specifically to this setting we are able to avoid this statistical penalty and obtain near-optimal computationally tractable robust estimators.

The remainder of the paper is organized as follows. In Section 2, we provide some background on robust estimation and formally introduce the examples we consider throughout this paper. Section 3 is devoted to our main results and their consequences. Section 4 includes a description of our main algorithm, and includes a sketch of its analysis with more technical details deferred to the Appendices. In Section 5 we provide improved algorithms for detection and estimation in sparse PCA. We conclude with a brief discussion of avenues for future work.

2. Background and Problem Setup

In this section we provide some background on robust estimation, before providing a precise definition of the statistical models we consider.

2.1. Robust estimation

In the robust estimation framework we suppose that we obtain samples $\{x_1, \ldots, x_n\}$ where each sample x_i is distributed according to the mixture model P_ϵ in Eqn. (1). In this model, the distribution Q is allowed to be completely arbitrary and represents the distribution of "outliers". An alternative viewpoint arises from the observation that the set of possible distributions P_ϵ is equivalent to the ℓ_1 ball around P of radius ϵ. Indeed, we can alternatively view desirable estimators in this model as those that are robust to model-misspecification (in the ℓ_1 or total variation metric).

Our focus, will be on finite-dimensional functionals of the target distribution P. Formally, for a given function $g : \mathbb{R}^{\widetilde{d}} \mapsto \mathbb{R}^d$, we define the corresponding functional $\theta_g : P \mapsto \mathbb{R}^d$, where:

$$\theta_g(P) = \mathbb{E}_{x \sim P}[g(x)].$$

Motivated by similar considerations in high-dimensional statistics, our sparsity assumption will be that the number of non-zeros $\|\theta_g(P)\|_0 \leq k$. We will further denote the covariance as,

$$\text{cov}(\theta_g(P)) = \mathbb{E}_{x \sim P}\left[(g(x) - \theta_g(P))(g(x) - \theta_g(P))^T\right] \tag{2}$$

Our algorithm will be based on appropriately weighting samples in order to match second order information and in order to accomplish this we will rely on the existence of a certain algebraic form for the covariance. In particular, we will suppose that there exists a multivariate function $F : \mathbb{R}^d \mapsto \mathbb{R}^{d \times d}$, such that $F(\theta_g(P)) = \text{cov}(\theta_g(P))$. An important restriction on this algebraic form, that will enable accurate estimation, is that it be sufficiently regular. We first assume

$$L_{\text{cov}} = \max_{\|v\|_2, \|v\|_0 \leq k} \left| v^\top \text{cov}\left(\theta_g\right) v \right| \tag{3}$$

for some constant L_{cov}. Second, we require that for any two vectors $\theta_1, \theta_2 \in \mathbb{R}^d$, there exist a constant L_F and a universal constant C such that

$$\|F(\theta_1) - F(\theta_2)\|_{\text{op}} \leq L_F \|\theta_1 - \theta_2\|_2 + C\|\theta_1 - \theta_2\|_2^2. \tag{4}$$

Our bounds depend explicitly on L_F and L_{cov}. In the next subsection, we consider a variety of examples and describe the appropriate functionals of interest and their corresponding covariances.

2.2. Illustrative examples

Our general results apply to a variety of statistical models and in this section we describe a few concrete examples of interest.

Sparse Gaussian Mean Estimation: In this setting, we observe samples

$$\{x_1, \ldots, x_n\} \sim (1 - \epsilon)N(\mu, I) + \epsilon Q, \tag{5}$$

where each $x_i \in \mathbb{R}^d$ and for an arbitrary Q [1]. The goal in this setting is to estimate μ in the ℓ_2 norm in the high-dimensional setting, under the assumption of sparsity, i.e. that $\|\mu\|_0 \leq k$. Using the notation introduced earlier, the function g is simply the identity, i.e. $g(x) = x$.

Sparse Gaussian Covariance Estimation: In this case, we observe samples

$$\{x_1, \ldots, x_n\} \sim (1 - \epsilon)N(0, \Sigma) + \epsilon Q, \tag{6}$$

where each $x_i \in \mathbb{R}^d$ and where the covariance matrix can be written as $\Sigma = I + \Omega$, where $\|\Omega\|_0 \leq k$. The goal in this problem is to estimate the sparse matrix Ω. This problem is closely related to the problem of Gaussian graphical modeling. Zeros in Σ signal marginal independencies, and this can be used to construct a graphical display of the relationship between the features (Bien and Tibshirani, 2011). In this problem, denoting by $\text{vec}(M)$ the vectorization of the matrix M, and by $\text{diag}(M)$ its diagonal entries, we consider the function $g(x) = \text{vec}(xx^T - \text{diag}(xx^T))$. Further, using $\otimes$ to denote the Kronecker product, we have that: $F(\text{vec}(S)) = \text{vec}(S)\text{vec}(S)^T + S \otimes S$.

Finally, we note that via a simple reduction scheme (described in detail in Diakonikolas et al. (2016a)) we can combine the above two settings in order to jointly estimate an unknown mean, and an unknown covariance robustly in a high-dimensional setting provided both are sparse.

Sparse PCA in the Spiked Covariance Model: A popular model in which to study issues that arise in the detection and estimation of high-dimensional principal components is the *spiked covariance model* introduced by (Johnstone, 2001). Here we obtain samples

$$\{x_1, \ldots, x_n\} \sim (1 - \epsilon)N(0, I + \rho vv^T) + \epsilon Q, \tag{7}$$

where v is a k-sparse vector, i.e. $\|v\|_0 = k$, $\|v\|_2 = 1$, ρ is the signal-to-noise ratio, and Q is arbitrary. In other words, the sparse matrix S in the sparse Gaussian covariance estimation problem is assumed to be rank 1. We focus on two canonical tasks in this setting:

[1] We address the unknown covariance case in the sequel.

1. The *detection* problem which is to test for the existence of the rank 1 spike. More formally, for some critical radius ρ_{crit} we consider distinguishing the two hypotheses:

$$H_0 : \rho = 0 \quad \text{and} \quad H_1 : \rho \geq \rho_{\text{crit}}. \tag{8}$$

   For any test $\psi : \{x_1, \ldots, x_n\} \mapsto \{0, 1\}$, we evaluate its detection risk, which is simply the sum of its Type I and (maximal) Type II errors, i.e.

$$R(\psi) = P_0(\psi = 1) + \sup_{\rho \geq \rho_{\text{crit}}} P_\rho(\psi = 0),$$

   where P_ρ corresponds to the distribution of samples in Eqn. (7), and P_0 corresponds to the same distribution with $\rho = 0$.

2. The *recovery* or estimation problem which is to estimate the (uncorrupted) covariance matrix in the Frobenius norm. This is equivalent to finding an estimate of the principal direction $\widehat{v}$, where our loss is measured as $L(\widehat{v}, v) := \frac{1}{\sqrt{2}} \left\vert\!\left\vert\!\left\vert \widehat{v}\widehat{v}^T - vv^T \right\vert\!\right\vert\!\right\vert_{\text{op}}$.

In the uncorrupted setting, it is folklore (see e.g. (Berthet and Rigollet, 2013a)) that $n = \Theta(k \log d/\rho^2)$ samples are necessary for detection, and moreover that the minimax rate for estimation is $\Theta(\sqrt{k \log d/n\rho^2})$ (Wang et al., 2016). However, for computationally efficient algorithms the best known results are that $n = O(k^2 \log d/\rho^2)$ suffice for detection, and to construct a tractable estimator whose error scales as $O(\sqrt{k^2 \log d/n\rho^2})$. Furthermore, this rate is tight assuming the planted clique hypothesis (Berthet and Rigollet (2013a); Wang et al. (2016)).

Linear Regression: Linear regression is a canonical problem in statistics. In the uncontaminated setting we observe paired samples $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ which are related via the linear model,

$$y_i = \langle x_i, \beta \rangle + \epsilon_i, \tag{9}$$

where $x_i, \beta \in \mathbb{R}^d$ and $\epsilon_i \in \mathbb{R}$ is some type of observation noise. In this paper, we assume that $x_i \sim N(0, I)$ and $\epsilon_i \sim N(0, 1)$, and our goal is to estimate the unknown β in a high-dimensional setting under the assumption that $\|\beta\|_0 \leq k$. In this problem, we take $g((y, x)) = yx$, by making the observation that the functional of interest $\beta = \mathbb{E}[yx]$. Further, we can calculate the algebraic form for the covariance as $F(\beta) = (\|\beta\|_2^2 + 1)I + \beta\beta^T$.

Generalized Linear Models (GLMs): We consider two distinct forms for GLMs in our work. The first form is a non-linear regression model where the uncontaminated distribution P corresponds to pairs $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ which are related as,

$$y_i = u(\langle x_i, \beta \rangle) + \epsilon_i \tag{10}$$

where u is a known non-linear function, $x_i, \beta \in \mathbb{R}^d$ and $\epsilon_i \in \mathbb{R}$. As before we assume that, $x_i \sim N(0, I)$, $\epsilon_i \sim N(0, 1)$, and further that there exist constants C_1 and C_2 such that, $u(0) \leq C_1$ and u is C_2-Lipschitz, i.e. for any pair $x, y \in \mathbb{R}$ we have that,

$$|u(x) - u(y)| \leq C_2 |x - y|.$$

The goal is to estimate the unknown, sparse β. In this case, we choose $g((y, x)) = \frac{xy}{\mathbb{E}[\nabla_{x'} u(x')]}$ where $x' = \langle x, \beta \rangle$. As a consequence of Stein's identity we have that $\mathbb{E}[g((y, x))] = \beta$. Once again, by Stein's identity (see Appendix E) we obtain the algebraic form of the covariance:

$$F(\beta) = \left(\frac{1 + \mathbb{E}[u^2(x')]}{(\mathbb{E}[\nabla_{x'} u(x')])^2} \right) I + \left(\frac{\mathbb{E}[2u(x')\nabla_x^2 u(x') + (\nabla u(x'))^2]}{(\mathbb{E}[\nabla'_x u(x')])^2} \right) \beta\beta^T.$$

where $x' = \langle x,\, \beta \rangle$. Observe that $F(\beta)$ has the form $\kappa_1 I + \kappa_2 \beta\beta^\top$ where κ_1 and κ_2 are scalars. Further notice that $x' \sim N\left(0, \|\beta\|_2^2\right)$, so these quantities can be estimated easily using just $\{y_1, \ldots, y_n\}$ with a one-dimensional robust method like the median estimator. Therefore, from now on, we will assume these quantities are known constants.

Logistic-type Models: Finally, our theory also applies to GLMs of the logistic regression form. In the uncontaminated setting, we observe pairs $\{(y_1, x_1), \ldots, (y_n, x_n)\}$, where $y_i \in \{0, 1\}$ where,

$$\mathbb{P}(y_i = 1 | x_i) = u(\langle x_i,\, \beta \rangle),$$

and the assumptions on x and u are as before. In this case, the function g is identical to the previous case, and its corresponding covariance is given as (see Appendix F):

$$F(\beta) = \left(\frac{\mathbb{E}[u(x')]}{(\mathbb{E}[\nabla_{x'} u(x')])^2} \right) I + \left(\frac{\mathbb{E}[\nabla^2_{x'} u(x') - (\nabla u(x'))^2]}{(\mathbb{E}[\nabla_{x'} u(x')])^2} \right) \beta\beta^T.$$

where $x' = \langle x,\, \beta \rangle$. With these preliminaries in place, we devote our next section to a description of our main results concerning the robust high-dimensional estimation of these statistical models.

3. Main Results

In this paper, our contributions are two-fold. Our first result is to provide a general framework for high dimensional robust recovery, which we show is powerful enough to encode many problems involving sparsity. This consists of two ingredients: (1) a set of deterministic conditions under which such recovery is possible, and then (2) concentration results which show that these deterministic conditions hold after not too many samples. This yields our results for sparse mean estimation, sparse covariance estimation, and sparse regression.

Our second result is to show that in the spiked covariance model, it is possible to move past this framework (which fails to provide tight bounds in this setting), and construct a new convex program for robust recovery. We first give a simple algorithm for the detection problem, then show that this algorithm can be modified somewhat to also yield recovery. Here, our rates, unlike elsewhere in the paper, match the computationally efficient rates without noise.

3.1. Notation

In this model defined in Eqn. (1), we can define two subsets of $\mathcal{G}, \mathcal{B} \subseteq \{1, \ldots, n\}$, where $i \in \mathcal{G}$ if the corresponding sample is drawn from P, and $i \in \mathcal{B}$ otherwise. Following (Diakonikolas et al., 2016a), we define a set of feasible weights as

$$S_{n,\epsilon} = \left\{ \{w_1, \ldots, w_n\} : \sum_{i=1}^n w_i = 1, 0 \le w_i \le \frac{1}{(1 - 2\epsilon)n} \;\forall\, i \right\}. \tag{11}$$

Noting that with high-probability there are fewer than $2\epsilon n$ points in the set $\mathcal{B}$, the set $S_{n,\epsilon}$ with high-probability contains the ideal weights which we denote w^* whose entries are given as,

$$w_i^* = \frac{\mathbb{I}(i \in \mathcal{G})}{|\mathcal{G}|} \quad \forall\, i. \tag{12}$$

For any given weight vector w, we define its renormalized restriction to the points in $\mathcal{G}$ and $\mathcal{B}$ via,

$$w_i^g = \frac{w_i}{\sum_{i \in \mathcal{G}} w_i} \triangleq \frac{w_i}{w_g} \quad w_i^b = \frac{w_i}{\sum_{i \in \mathcal{B}} w_i} \triangleq \frac{w_i}{w_b} \quad \forall\, i.$$

With this notation in place, we can further define a collection of quantities of interest for a fixed set of weights $w \in S_{n,\epsilon}$. A naive estimator of the functional is simply

$$\widetilde{\theta}(w) = \sum_{i=1}^{n} w_i g(x_i), \tag{13}$$

and its error is denoted as $\widetilde{\Delta}(w) = \widetilde{\theta}(w) - \theta_g(P)$. A more nuanced estimator further exploits the expected sparsity of the functional by truncating its smaller entries. We define, for a positive vector v, $P_k(v)$ to be the vector where the k-th largest entries in magnitude are retained (breaking ties arbitrarily) and all other entries are set to 0. Then we define,

$$\widehat{\theta}(w) = P_{2k}(\widetilde{w}) \tag{14}$$

and its error $\widehat{\Delta}\left(w\right) = \widehat{\theta}\left(w\right) - \theta_g\left(P\right)$. Recalling, the definition of the covariance functional in Eqn. (2) we define the error of the weighted covariance as,

$$\mathcal{E}(w) = \sum_{i=1}^{n} w_i (g(x_i) - \theta_g(P))(g(x_i) - \theta_g(P))^T - \text{cov}(\theta_g(P)).$$

In allowing for a high-dimensional scaling, where $d \gg n$, we can no longer expect $\widetilde{\Delta}(w)$ to be small in an ℓ_2 sense and $\mathcal{E}(w)$ to be small in an operator norm sense. Instead, we rely on establishing a more limited control on these quantities. We define the k-sparse operator norm as,

$$\left\lvert\!\left\lvert\!\left\lvert M \right\rvert\!\right\rvert\!\right\rvert_{\text{k,op}} = \max_{S \subset [d], |S| \leq k} \left\lvert\!\left\lvert\!\left\lvert M^{SS} \right\rvert\!\right\rvert\!\right\rvert_{\text{op}}.$$

Finally, we define $\|M\|_\infty = \max_{i,j} |M_{ij}|$.

3.2. A general framework for robust recovery

With these definitions in place we can now state our main deterministic result. We begin by identifying a set of deterministic conditions under which we can design a polynomial time algorithm that is provably robust. We focus on functionals θ_g for which Equations (3) and (4) are satisfied. Our main deterministic result is the following:

Theorem 3.1 (Main Theorem) *Suppose that, for samples $\{x_1, \ldots, x_n\}$ drawn from the ϵ-contamination model, we have that $\|\theta_g(x_i)\|_2 \leq D$, and further that there exist a universal constant C_1 such that* *the following conditions hold:*

$$
\begin{aligned}
|\mathcal{B}| &\leq 2\epsilon n, && (15)\\
\|\widetilde{\Delta}(w^*)\|_\infty &\leq C_1\left(\frac{\left(L_F+\sqrt{L_{\text{cov}}}\right)\delta}{k}\right), && (16)\\
\|P_k(\widetilde{\Delta}(w^g))\|_2 &\leq C_1\left(\left(L_F+\sqrt{L_{\text{cov}}}\right)\delta\right) \;\; \forall\, w\, \in S_{n,\epsilon}, && (17)\\
\|\mathcal{E}(w^*)\|_\infty &\leq C_1\left(\frac{\left(L_F^2+L_{\text{cov}}\right)\delta}{k}\right), && (18)\\
\left\|\!\left|\mathcal{E}(w^g)\right|\!\right\|_{k,op} &\leq C_1\left(\left(L_F^2+L_{\text{cov}}\right)\delta\right) \;\; \forall\, w\, \in S_{n,\epsilon} && (19)
\end{aligned}
$$

for some $\delta = \Omega(\epsilon)$. Then there is an algorithm which runs in time polynomial in $\left(n, d, \frac{1}{\epsilon}\right)$ and outputs $\widehat{\theta}$ satisfying $\|\widehat{\theta}-\theta\|_2 \leq C_2\left(\left(\sqrt{L_{\text{cov}}}+L_F\right)\delta\right)$ for some absolute constant C_2.

Several remarks are in order. In order to apply the theorem to a specific statistical model, we simply need to verify that the functional is sufficiently regular (see Equations (3) and (4)), that the functional is bounded by a polynomial in $(n, d, D, 1/\epsilon)$, and finally that the conditions in Equations (15)-(19) are satisfied. We ensure boundedness via a simple pruning step that removes gross, and easily detectable, outliers. In order to verify the main deviation conditions of the theorem, we note that there are two types of deviation we need to control. The first type in Equations (16) and (18) establishes strong ℓ_∞ control, decaying with the sparsity k, but only needs to hold for the ideal weights w^*. The other type of control, in Equations (17) and (19) is on an k-sparse operator norm and needs to hold *uniformly* over the set $S_{n,\epsilon}$, but importantly ignores the weights on the points in $\mathcal{B}$ via restriction to w^g. In concrete examples, we establish the latter control via the use of empirical process arguments (selecting an appropriate covering and using the union bound).

Re-visiting Illustrative Examples We now turn our attention to the statistical problems introduced earlier, and derive specific corollaries of our deterministic result. We begin with the case of estimating a sparse Gaussian mean, when the covariance is the identity.

Corollary 3.1 (Robust Estimation of Sparse Gaussian Mean) *Consider the model introduced in Equation (5), then there are universal constants C_1, C_2 such that, if $n \geq C_1\left(\frac{k^2\log(d/\tau)}{\epsilon^2\log 1/\epsilon}\right)$, then there exists an algorithm that runs in time polynomial in (d, n) and outputs an estimate $\widehat{\mu}$ that with probability at least $1-\tau$ satisfies: $\|\widehat{\mu}-\mu\|_2^2 \leq C_2\epsilon^2\log\frac{1}{\epsilon}$.*

It is worth noting that in contrast to prior work the sample complexity, has a logarithmic dependence on the ambient dimension d, allowing for high-dimensional scalings where $d \gg n$, provided that the sparsity $k^2 \ll n$. As in the work of Diakonikolas et al. (2016a), we obtain near-optimal contamination dependence scaling upto a logarithmic factor as roughly ϵ^2. Importantly, as emphasized in prior work (Diakonikolas et al., 2016a; Lai et al., 2016) and in stark contrast to other tractable robust estimators, the contamination dependence achieved by our algorithm is completely independent of the dimension of the problem. In comparing to information-theoretic lower bounds (see Appendix B), we notice that the sample complexity is worse by a factor k. As will be clearer in the sequel, this increased sample complexity is due to use of a convex relaxation for sparse PCA (d'Aspremont et al., 2007). This phenomenon, arises in a variety of statistical estimation problems and is believed to be related to the hardness of the planted clique problem (Berthet and Rigollet, 2013b). Next, we consider the performance of our method, in estimating a sparse covariance matrix.

Corollary 3.2 (Robust Sparse Gaussian Covariance Estimation) *Consider the model introduced in Equation (6). There are universal constants C_1, C_2 such that if the sample size $n \geq C_1 \left(\frac{k^2 \log(d/\tau)}{\epsilon^2}\right)$, then there is an algorithm that runs in time polynomial in (d, n) and produces an estimate $\widehat{\Omega}$ that with probability at least $1-\tau$ satisfies: $\left\|\left|\widehat{\Omega} - \Omega\right\|\right|_F^2 \leq C_2 \left(\|\|\Omega\|\|_F^2 \epsilon^2 \log^4 \frac{1}{\epsilon}\right)$.*

We note that once again, the result is applicable even when $n \ll d$, that the statistical estimation rate is optimal up to a factor of k and that the contamination dependence is optimal up to logarithmic factors. Lastly, we apply our estimator to the various generalized linear models introduced earlier.

Corollary 3.3 (Robust Sparse Generalized Linear Models) *Consider the models in Equations (9),(10), and (11). If the target parameter β satisfies, $\|\beta\|_2 \leq \rho$, then there exist universal constants C_1, C_2 such that if $n \geq C_1 \left(\frac{k^2 \log(d/\tau)}{\epsilon^2}\right)$, then there exists an algorithm that runs in time polynomial in (d, n, ρ) and produces an estimate $\widehat{\beta}$ such that with probability at least $1-\tau$:*

1. *Linear and Generalized Linear Models: $\|\widehat{\beta} - \beta\|_2^2 \leq C_2 \left(\left(\|\beta\|_2^2 + 1\right) \epsilon^2 \log^4 \frac{1}{\epsilon}\right)$.*

2. *Logistic-type Models: $\|\widehat{\beta} - \beta\|_2^2 \leq C_2 \left(\left(\|\beta\|_2^2 + 1\right) \epsilon^2 \log^2 \frac{1}{\epsilon}\right)$.*

By exploiting the natural boundedness of the logistic-type models, we are able to obtain slightly stronger guarantees than in the regression setting.

3.3. The spiked covariance model for sparse PCA

We now turn our attention to the problems of robust detection and estimation in the spiked covariance model. Recall in this model, samples are generated according to the model in Eqn. (7). Observe that if we simply apply Corollary 3.2, then since vv^T is k^2 sparse, we obtain a rate of $O(k^4 \log d/\epsilon^2)$ for recovery—which is off by a factor of k^2 from the optimal rate. We develop a tailored approach for the sparse PCA setting which yields a faster (near-optimal) rate. Focusing first on the detection problem, we fix $\delta, \epsilon > 0$ and for a sufficiently small universal constant $c > 0$ we define the critical radius as: $\rho_{\text{crit}} = c\sqrt{\frac{\min(d,k^2) + \log\binom{d^2}{k^2} + \log 1/\delta}{n}}$. We obtain the following result:

Theorem 3.2 (Robust sparse PCA detection) *Suppose that, $\epsilon\sqrt{\log(1/\epsilon)} = O(\rho_{crit})$, and consider the hypothesis testing problem in Eqn. (8). Then Algorithm 3 has detection risk at most δ.*

We then modify our algorithm to give an analogous rate for recovery:

Theorem 3.3 (Robust sparse PCA recovery) *Fix $\epsilon, \rho > 0$, and consider the sparse PCA estimation problem. There is a universal constant $C > 0$ such that if $n \geq C\left(\frac{\min(d,k^2) + \log\binom{d^2}{k^2} + \log 1/\delta}{\rho^2}\right)$, then there is an efficient algorithm, which with probability at least $1-\delta$ outputs a vector $\widehat{v}$ such that*

$$L(\widehat{v}, v) = O\left(\frac{(1+\rho)\epsilon\sqrt{\log(1/\epsilon)}}{\rho}\right).$$

Algorithm 1 Robust Sparse Functional Estimation

```
Input: $\{x_1, \ldots, x_n\}$, $\tau_{\text{prune}}$, $k$, $\tau_{\text{sep}}$
Run a naive pruning algorithm, with input $(\{x_1, \ldots, x_n\}, \tau_{\text{prune}})$ and output $\{z_1, \ldots, z_m\}$.
Run the ellipsoid algorithm using the separation oracle described in Algorithm 2 with input $(\{z_1, \ldots, z_m\}, s, \tau_{\text{sep}})$ and output $\{w_1, \ldots, w_m\}$.
Output: $\widehat{\theta} = P_{2k}\left(\sum_{i=1}^m w_i g(z_i)\right)$.
```

4. A Unified Algorithm for Robust Sparse Functional Estimation

Broadly, our main algorithm follows the template of the convex programming approach for Gaussian mean estimation in Diakonikolas et al. (2016a), described in Algorithm 1. The algorithm proceeds in two steps, first a naive pruning step is applied to remove clear outliers in order to ensure that various quantities remain bounded by a radius that is polynomial in $(n, d, 1/\epsilon)$. In the sequel, we use $\{z_1, \ldots, z_m\}$ to denote the pruned sample. We generalize a similar pruning step from prior works (Diakonikolas et al., 2016a; Lai et al., 2016) to deal with the generalized linear model settings. This in turn further ensures that the subsequent use of the ellipsoid algorithm, terminates in polynomial time. At a high-level the ellipsoid algorithm is used to exploit the covariance structure of the functional in order to obtain a weighting of the sample that appropriately down-weighs detrimental samples from the contamination distribution Q.

Separation oracle via sparse PCA: Our first main technical contribution, is a new separation oracle, appropriate for the high-dimensional setting. The separation oracle in the work of Diakonikolas et al. (2016a) is based on the operator norm deviation between the weighted empirical covariance from its known or anticipated form. In the high-dimensional setting when $n \ll d$, even in the absence of outliers the covariance function cannot be estimated well in the operator norm. Exploiting the sparsity of the underlying functional we show that it suffices instead to ensure that the weighted empirical covariance is close to its anticipated form only on k-sparse subsets of the coordinates. However, this leads to the next technical hurdle: we need to be able to detect the deviation of the weighted empirical covariance on sparse subsets. This is the sparse PCA problem and is known to be NP-hard in a strong sense (Tillmann and Pfetsch, 2014). We consider instead using a convex relaxation for sparse PCA (d'Aspremont et al., 2007), and show upto a loss of a factor of k in the sample complexity, this convex relaxation suffices to construct our separation oracle.

Hard-thresholding with redundancy: Even in the absence of outliers, the natural estimator for a functional – its empirical counterpart – is inconsistent when $n \ll d$, at least in an ℓ_2 sense. However, the empirical estimator remains adequate both in an ℓ_∞ sense, and over sparse subsets. In order to exploit this insight when the true functional is sparse we use a careful truncation at various points in order to establish appropriate error control. A key aspect of this truncation is to ensure a certain redundancy by retaining roughly twice as many entries at each step, which allows us to adequately control the possible bias induced by truncation.

General forms for the covariance: A final conceptual contribution that we highlight is generalizing the basic insight of the work of Diakonikolas et al. (2016a). At a high-level, a key observation of their work is that in cases where the covariance structure is either known or in some sense related to the mean structure, this fact can be exploited in order to identify good weightings of the samples. We generalize this insight, identifying a set of smoothness conditions on the covariance map (see

Algorithm 2 Separation Oracle for Robust Sparse Estimation

```
Input: Weights from the previous iteration $\{w_1, \ldots, w_m\}$, pruned samples $\{z_1, \ldots, z_m\}$, tolerance parameter $\tau_{\text{sep}}$, sparsity level $k$.
Compute $\widehat{\theta} = P_{2k}\left(\sum_{i=1}^m w_i g\left(z_i\right)\right)$ and $E = \sum_{i=1}^m w_i \left(g\left(z_i\right) - \widehat{\theta}\right) \otimes \left(g\left(z_i\right) - \widehat{\theta}\right) - F(\widehat{\theta})$.
Solve the following convex program [2]:
    $$\max_H \quad \mathrm{tr}\left(EH\right) \quad \text{subject to } H \succcurlyeq 0 \quad \left\|H\right\|_{1,1} \leq k \quad \mathrm{tr}\left(H\right) = 1. \qquad (20)$$
    Let $H^*$ be the solution and $\lambda^*$ be the optimal value.
if $\lambda^* \leq \tau_{\text{sep}}$ then
    Return: "Yes".
else
    Return: The separating hyperplane:
    $$\ell(w') = \mathrm{tr}\Big(\Big[\Big(\sum_{i=1}^m w_i'\big(g\left(z_i\right) - \widehat{\theta}\big) \otimes \big(g\left(z_i\right) - \widehat{\theta}\big)\Big] - F\big(\widehat{\theta}\big)\Big)H^*\Big) - \lambda^*.$$
end if
```

Equations (3) and (4)) that allow us to tractably exploit the covariance structure. Concretely, deriving the covariance structure for mean and covariance estimation, GLMs and logistic-type models and showing that they satisfy these conditions enables a unified treatment.

5. Algorithms for Robust Sparse PCA

We first describe our techniques at a high level. Our main idea is to treat the SDP for sparse PCA as a norm, and then to write convex optimization problems over the set $S_{n,\epsilon}$ directly using that norm. Formally, for any convex set $S \subseteq \mathbb{R}^{d\times d}$, let $\left\|M\right\|_S^* = \sup_{A\in S} \left|\mathrm{tr}(AM)\right|$ denote the dual norm induced by S. In particular, observe that if we let

$$\mathcal{X}_k = \{H : H \succeq 0, \left\|H\right\|_{1,1} \leq k, \mathrm{tr}(H) = 1\},$$

then $\left\|M\right\|_{\mathcal{X}_k}^*$ is exactly the absolute value of the solution to (20), the SDP relaxation for sparse PCA. This allows us to write down optimization of $\left\|\cdot\right\|_S^*$ as an SDP which can be solved efficiently.

We now concretely describe how this can be used in robust sparse PCA . For detection, we show that, as before, if we can find weights on the samples so that the empirical covariance with these samples has minimal dual norm, then the value of the dual norm gives us a distinguisher between the spiked and non-spiked case. To find such a set of weights, we observe that norms are convex, and thus our objective is convex. Thus, as before, to optimize over this set it suffices to give a separation oracle, which the SDP for sparse PCA provides us.

We now turn our attention to the recovery problem. Here, the setup is very similar, except now we simultaneously find a set of weights and an "explainer" matrix A so that the empirical covariance with these weights is "maximally explained" by A, in a norm very similar to the one induced by the sparse PCA SDP. Utilizing that norms are convex, we show that this can be done via a convex program using the types of techniques described above, and that the top eigenvector of the optimal

A gives us the desired solution. While the convex program would be quite difficult to write down in one shot, it is quite easily expressible using the abstraction of dual norms.

5.1. The detection problem

In this section, we give an efficient algorithm for detecting a spiked covariance matrix in the presence of adversarial noise. Our algorithm is fairly straightforward: we ask for the set of weights $w \in S_{n,\epsilon}$ so that the empirical second moment with these weights has minimal deviation from the identity in the dual $\mathcal{X}_k$ norm. We may write this as a convex program. Then, we check the value of the optimal solution of this convex program. If this value is large, then we reject the null hypothesis, i.e. return the value 1 and otherwise we return the value 0. The formal description of this algorithm is given in Algorithm 3. From this description it is not immediately clear that this algorithm can be implemented efficiently. In Appendix G.1 we show this is not an issue. At the heart of the matter is that $\left|\!\left|\!\left| \cdot \right|\!\right|\!\right|_{\mathcal{X}_k}$ can be computed efficiently, and that the specific $A \in \mathcal{X}_k$ which achieves the maxima can also be found efficiently.

Algorithm 3 Robust detection of a rank 1 spike

```
1: Input: samples $x_1, \ldots, x_n$, error parameter $\epsilon$, failure parameter $\delta$, signal to noise ratio $\rho$
2: Let $\gamma$ be the value of the solution
       $$\min_{w \in S_{n,\epsilon}} \left|\!\left|\!\left| \sum_{i=1}^n w_i (x_i x_i^T - I) \right|\!\right|\!\right|_{\mathcal{X}_k}^* \qquad (21)$$
3: if $\gamma \geq \rho/2$ then reject the null hypothesis (return 1), else accept (return 0).
```

5.2. The recovery problem

We now describe how to solve the robust recovery problem. An initial attempt would be to simply run the same SDP in (21), and hope that the dual norm maximizer gives you enough information to recover the hidden spike. This would more or less correspond to the simplest modification SDP of the sparse PCA in the non-robust setting that one could hope gives non-trivial information in this setting. However, this cannot work, for the following straightforward reason: the value of the SDP is always at least $O(\rho)$, as we argue in Section G.2. Therefore, the noise can pretend to be some other sparse vector u orthogonal to v, so that the covariance with noise looks like $w^g(I+\rho vv^T)+w^g\rho uu^T$, so that the value of the SDP can be minimized with the uniform set of weights. Then it is easily verified that both vv^T and uu^T are dual norm maximizers, and so the dual norm maximizer does not uniquely determine v.

To circumvent this, we simply add a slack variable to the SDP, which is an additional matrix in $\mathcal{X}_k$, which we use to try to maximally explain away the rank-one part of $I + \rho vv^T$. This forces the value of the SDP to be small, which allows us to show that the slack variable actually captures v.

Our algorithms and analyses will make crucial use of the following convex set, which is a relaxation of $\mathcal{X}_k$:

$$\mathcal{W}_k = \left\{ X \in \mathbb{R}^{d \times d} : \operatorname{tr}(X) \leq 1, \|X\|_{1,1} \leq k, X \succeq 0 \right\} .$$

Our algorithm, given in Algorithm 4, is the following. We solve a convex program which simultaneously chooses a weights in $S_{n,\epsilon}$ and a matrix $A \in \mathcal{W}_k$ to minimize the $\mathcal{W}_k$ distance between the sample covariance with these weights, and A. Our output is the top eigenvector of A.

Algorithm 4 Robust estimation of the top principal component

```
Input: samples $x_1, \ldots, x_n$, error rate $\epsilon$, failure probability $\delta$, signal to noise ratio $\rho$
Let $w^*$, $A^*$ be the solution to
    $\mathrm{argmin}_{w \in S_{n,\epsilon}, A \in \mathcal{X}_k} \left\| \sum_{i=1}^{n} w_i (x_i x_i^T - I) - \rho A \right\|_{\mathcal{W}_{2k}}^{*}$    (22)
Let $u$ be the top eigenector of $A^*$
return $P_k(v)$
```

6. Conclusion and Future Directions

In this paper we propose a computationally tractable robust algorithm for sparse high-dimensional statistical estimation problems. We develop a general result, which we then specialize to obtain corollaries for sparse mean/covariance estimation, sparse linear regression and sparse generalized linear models. In each of these problems, we obtain near optimal dependency on the contamination parameter, and sample complexities that depend only logarithmically on the ambient dimension.

Future directions of research include developing faster alternatives to the ellipsoid algorithm, to further relax the assumptions in various settings, and finally to close the gap in sample complexity to statistically optimal, albeit computationally intractable procedures (Chen et al., 2015, 2016).

Acknowledgments

S.B. was supported by the NSF grant DMS-1713003. S.S.D was supported by ARPA-E under grant in the the Terra program (Transportation Energy Resources from Renewable Agriculture) with title Automated TERRA Phenotyping System for Genetic Improvement of Energy Crops. J.L. was supported by NSF grant CCF-1217921, DOE grant DE-SC0008923, NSF CAREER Award CCF-145326, and a NSF Graduate Research Fellowship. The authors would like to thank Michael Cohen, Ilias Diakonikolas, Ankur Moitra, Ludwig Schmidt and Martin Wainwright for useful discussions. A.S. was supported by the NSF CAREER grant IIS-1252412 and CCF-1563918.

References

Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.

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

Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*, 2013b.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4): 807, 2011.

Siu-On Chan, Ilias Diakonikolas, Xiaorui Sun, and Rocco A Servedio. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1380–1394. SIAM, 2013.

Siu On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *Advances in Neural Information Processing Systems*, pages 1844–1852, 2014.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. *arXiv preprint arXiv:1611.02315*, 2016.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*, 2015.

Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for huber's $epsilon$-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning k-modal distributions via testing. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1371–1385. Society for Industrial and Applied Mathematics, 2012.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. *arXiv preprint arXiv:1604.06443*, 2016a.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint arXiv:1606.03077*, 2016b.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. *arXiv preprint arXiv:1611.03473*, 2016c.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

Cecil Hastings Jr, Frederick Mosteller, John W Tukey, and Charles P Winsor. Low moments for small samples: a comparative study of order statistics. *The Annals of Mathematical Statistics*, pages 413–426, 1947.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.

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

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 04 2001.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 1998.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. *arXiv preprint arXiv:1604.06968*, 2016.

Malik Magdon-Ismail. Np-hardness and inapproximability of sparse pca. *arXiv preprint arXiv:1502.05675*, 2015.

Michael L Overton and Robert S Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca for spiked random matrices and synchronization. *arXiv preprint arXiv:1609.05573*, 2016.

Charles Stein. Dependent random variables. 1971.

Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531, 1975.

Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in neural information processing systems*, pages 2670–2678, 2013.

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

Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Appendix A. Proofs from Section 4

We first give a high-level descriptions of key steps in our algorithm and corresponding lemmas.

Hard Thresholding: The idea of using hard thresholding in sparse estimation problems in order to ensure that the overall estimation error is well controlled, has been explored recently in iterative hard thresholding algorithms (see for instance Bhatia et al. (2015); Jain et al. (2014)). The key result we need, relates the k-sparse subset error of the original estimator to the full ℓ_2 error of the hard-thresholded estimator. Recalling the definitions of the error of the original estimator $\widetilde{\Delta}$ and the error of the thresholded estimator $\widehat{\Delta}$ we show the following result:

Lemma A.1 *Suppose θ_g is k-sparse, then we have the following result:*

$$\frac{1}{5}\|\widehat{\Delta}\|_2 \leq \left\|P_k\left(\widetilde{\Delta}\right)\right\|_2 \leq 4\|\widehat{\Delta}\|_2.$$

Intuitively, this result lets us pass from the high-dimensional feasible error control on subsets to the more desirable ℓ_2 error control.

Good Weights and Approximation of the Covariance: The utility of the ellipsoid algorithm is in finding an appropriate set of weights, such that the weighted empirical estimate of the functional is sufficiently accurate. In more details, we consider weights such that the weighted covariance is close to the true one on every sparse subset of coordinates. Defining, $\widehat{\theta} = P_{2k}\left(\sum_{i=1}^m w_i g\left(z_i\right)\right)$:

Definition A.1 *[Good Weights] Let C_δ be the subset of $S_{m,\epsilon}$ such that for any $w \in C_\delta$ we have*

$$\left\|\!\left|\sum_{i=1}^m w_i\big(g\left(z_i\right) - \widehat{\theta}\big)\big(g(z_i) - \widehat{\theta}\big)^\top - \mathrm{cov}\left(\theta_g\left(P\right)\right)\right|\!\right\|_{k,op} \leq \left(L_F^2 + L_{\mathrm{cov}}\right)\delta.$$

The parameter δ in the above definition is an accuracy parameter that will be chosen as a function of only ϵ differently for each model. The central role of this set of weights is captured by the following result, whose proof follows along similar lines to that of Lemma 4.19 of Diakonikolas et al. (2016a).

Lemma A.2 *Let $w \in S_{n,\epsilon}$ and suppose that for a universal constant C_1 we have,*

$$\begin{aligned}
|\mathcal{B}| &\leq 2\epsilon n, \\
\|P_s(\widetilde{\Delta}(w^g))\|_2 &\leq C_1\left(\left(L_F + \sqrt{L_{\mathrm{cov}}}\right)\delta\right), \\
|\!|\!|\mathcal{E}(w^g)|\!|\!|_{k,op} &\leq C_1\left(\left(L_F^2 + L_{\mathrm{cov}}\right)\delta\right),
\end{aligned}$$

where $\delta \geq C_2\epsilon$ for some sufficiently large constant C_2. If $\left\|P_s\left(\widetilde{\Delta}\left(w\right)\right)\right\|_2 \geq C_3\left(L_F + \sqrt{L_{\mathrm{cov}}}\right)\delta$ for some sufficiently large constant C_3, then for sufficiently small ϵ we have that,

$$\left\|\!\left|\sum_{i=1}^m w_i\big(g\left(z_i\right) - \widehat{\theta}\big)\big(g\left(z_i\right) - \widehat{\theta}\big)^\top - \mathrm{cov}\left(\theta_g(P)\right)\right|\!\right\|_{k,op} \geq \frac{\left\|P_k\left(\widetilde{\Delta}\left(w\right)\right)\right\|_2^2}{4\epsilon}.$$

Roughly, this lemma guarantees that if the weighting scheme is such that the error $\widetilde{\Delta}^S$ is large (in ℓ_2) then the weights cannot belong to the set of good weights defined above. We note that an essentially identical result can be proved if we replace the true covariance by a plug-in estimate, provided the covariance map is sufficiently smooth (see Lemma A.6). This results in an important reduction, in order to obtain an accurate estimate it suffices to find a weight vector that belongs to the set of good weights. We accomplish this via the ellipsoid algorithm.

Convex Relaxation of Sparse PCA: In order to use the previous lemma in the ellipsoid algorithm, we need to be able to design a separation oracle for the set of good weights. The main technical hurdle is that we need to compute, for a given set of weights, the sparse operator norm which is an intractable problem in general (Magdon-Ismail, 2015).

We replace the sparse PCA algorithm by a standard tractable convex relaxation (d'Aspremont et al., 2007). The following result shows that the optimal value of this program is sandwiched by the optimal value of the intractable sparse PCA program.

Lemma A.3 *For a fixed w, the optimal value $\lambda^*(w)$ of Eqn. (20) satisfies*

$$\lambda^*(w) \geq \left\| \sum_{i=1}^{m} w_i \big(g(z_i) - \widehat{\theta}(w)\big)\big(g(z_i) - \widehat{\theta}(w)\big)^\top - F\big(\widehat{\theta}(w)\big) \right\|_{k,op}.$$

Furthermore, the solution $H^*(w)$ satisfies that there is a universal constant C such that for any $w' \in S_{m,\epsilon}$

$$\begin{aligned}&\operatorname{tr}\left(\Big(\sum_{i=1}^{m} w_i'\big(g(z_i) - \widehat{\theta}(w)\big)\big(g(z_i) - \widehat{\theta}(w)\big) - F\big(\widehat{\theta}(w)\big)\Big)H^*(w)\right)\\ \leq & C\left(s\left\|\mathcal{E}(w')\right\|_\infty + \left\|\widehat{\Delta}(w)\right\|_2^2 + \left(L_F + s\left\|\widetilde{\Delta}(w')\right\|_\infty\right)\left\|\widehat{\Delta}(w)\right\|_2\right).\end{aligned}$$

Concretely, the above lemma provides two guarantees. First that the optimal value of the relaxation is never too small, so that the ellipsoid algorithm does not falsely accept a bad weighting scheme, and finally, that the separating hyperplane is sufficiently accurate when appropriate control can be established on the various stochastic fluctuations. We combine these two facts to complete the analysis of the ellipsoid algorithm and to establish Theorem 3.1 in the Appendix.

A.1. Proof of Hard Thresholding

Lemma A.4 (Lemma A.1) *Suppose θ_g is k-sparse, then we have the following result:*

$$\frac{1}{5}\|\widehat{\Delta}\|_2 \leq \left\|P_k\left(\widetilde{\Delta}\right)\right\|_2 \leq 4\|\widehat{\Delta}\|_2.$$

Proof We denote $\mathcal{S}^*$ to be the support for θ and $\mathcal{S}$ be indices for the selected $2k$ entries. We first prove $\frac{1}{5}\left\|\widehat{\Delta}\right\|_2 \leq \left\|P_k\left(\widetilde{\Delta}\right)\right\|_2$. Let $\tau = \left\|P_k\left(\widetilde{\Delta}\right)\right\|_2$. We have

$$\left\|\widehat{\theta} - \theta\right\|_2 \leq \left\|\widehat{\theta}^{\mathcal{S}} - \theta_g^{\mathcal{S}}\right\|_2 + \left\|\widehat{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)} - \theta_g^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 + \left\|\widehat{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)^c} - \theta_g^{\mathcal{S}^c\cap(\mathcal{S}^*)^c}\right\|_2.$$

Now we bound the three terms in the right hand side separately. The first term is bounded by 2τ by our assumption. The third term is 0 by definition of $\mathcal{S}$ and $\mathcal{S}^*$. For the second term, note

$$\begin{aligned}\left\|\widehat{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)} - \theta_g^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 &\leq \left\|\widetilde{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)} - \theta_g^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 + \left\|\widetilde{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)} - \widehat{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 \\ &= \left\|\widetilde{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)} - \theta_g^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 + \left\|\widetilde{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2.\end{aligned}$$

We have $\left\|\widetilde{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)} - \theta_g^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 \leq \tau$ by assumption. Assume $\left\|\widetilde{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 \geq 2\tau$. Since $|\mathcal{S}| = 2k$, then there exists $\mathcal{S}' \subset \mathcal{S}$, $|\mathcal{S}'| = k$ such that $\mathcal{S}^* \cap \mathcal{S}' = \emptyset$ with $\left\|\widetilde{\theta}^{\mathcal{S}'}\right\|_2 \geq \left\|\widetilde{\theta}^{\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2 \geq 2\tau$. However, $\left\|\widetilde{\theta}^{\mathcal{S}'}\right\|_2 = \left\|\widetilde{\Delta}^{\mathcal{S}'}\right\|_2 \leq \tau$ by our assumption. Therefore, $\left\|\widetilde{\theta}^{\mathcal{S}^c\cap(\mathcal{S}^*)}\right\|_2 \leq 2\tau$. Adding all these terms up, we have $\left\|\widehat{\theta} - \theta_g\right\|_2 \leq 4\tau$.

For the other direction, let $\gamma = \left\|\widehat{\Delta}\right\|_2$. For any $\mathcal{S}' \subset [d]$, $|\mathcal{S}'| \leq k$, we have

$$\begin{aligned}\left\|\widetilde{\Delta}^{\mathcal{S}'}\right\|_2 &\leq \left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}}\right\|_2 + \left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c}\right\|_2 \\ &= \left\|\widehat{\Delta}^{\mathcal{S}'\cap\mathcal{S}}\right\|_2 + \left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c}\right\|_2 \\ &\leq \gamma + \left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c}\right\|_2\end{aligned}$$

where the last inequality is by our assumption. Now applying triangle inequality on $\left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c}\right\|_2$, we have

$$\left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c}\right\|_2 \leq \left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2 + \left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c\cap(\mathcal{S}^*)^c}\right\|_2.$$

For the first term, observe that

$$\begin{aligned}\left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2 &\leq \left\|\widetilde{\Delta}^{\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2 \\ &= \left\|\widetilde{\theta}^{\mathcal{S}^c\cap\mathcal{S}^*} - \theta_g^{\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2 \\ &\leq \left\|\widetilde{\theta}^{\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2 + \left\|\theta_g^{\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2\end{aligned}$$

By definition of P_{2k}, there exists $\mathcal{S}'' \subset \mathcal{S}$ with $|\mathcal{S}''| = s$ and $\theta^{\mathcal{S}''} = 0$ and $\left\|\widetilde{\theta}^{\mathcal{S}''}\right\|_2 \geq \left\|\widetilde{\theta}^{\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2$. Therefore, $\left\|\widetilde{\theta}_{(\mathcal{S})^c\cap\mathcal{S}^*}\right\|_2 \leq \gamma$. Next, notice $\|\theta_g^{\mathcal{S}^c\cap\mathcal{S}^*}\|_2 = \left\|\widehat{\Delta}^{\mathcal{S}^c\cap\mathcal{S}^*}\right\|_2 \leq \gamma$. Lastly, note again $|\mathcal{S}' \cap (\mathcal{S})^c \cap (\mathcal{S}^*)^c| \leq s$, so

$$\begin{aligned}\left\|\widetilde{\Delta}^{\mathcal{S}'\cap\mathcal{S}^c\cap(\mathcal{S}^*)^c}\right\|_2 &= \left\|\widetilde{\theta}^{\mathcal{S}'\cap\mathcal{S}^c\cap(\mathcal{S}^*)^c}\right\|_2 \\ &\leq \left\|\widetilde{\theta}^{\mathcal{S}''}\right\|_2 \\ &= \left\|\widetilde{\Delta}^{\mathcal{S}''}\right\|_2 \\ &= \left\|\widehat{\Delta}^{\mathcal{S}''}\right\|_2 \\ &\leq \gamma.\end{aligned}$$

Therefore $\widetilde{\Delta}^{\mathcal{S}'} \leq 4\gamma$. Because $\mathcal{S}'$ is arbitrary, our proof is complete. ■

A.2. Proofs of Good Weights and Approximation of the Covariance

Lemma A.5 (Lemma A.2) *Let $w \in S_{m,\epsilon}$ and suppose that for a universal constant C_1 we have,*

$$\begin{aligned} |\mathcal{B}| &\leq 2\epsilon n, \\ \|P_s(\widetilde{\Delta}(w^g))\|_2 &\leq C_1 \left(\left(L_F + \sqrt{L_{\text{cov}}} \right) \delta \right), \\ \|\!|\mathcal{E}(w^g)|\!\|_{k,op} &\leq C_1 \left(\left(L_F^2 + L_{\text{cov}} \right) \delta \right), \end{aligned}$$

where $\delta \geq C_2 \epsilon$ for some sufficiently large constant C_2. If $\left\| P_k \left(\widetilde{\Delta}\left(w\right) \right) \right\|_2 \geq C_3 \left(L_F + \sqrt{L_{\text{cov}}} \right) \delta$ for some sufficiently large constant C_3, then for sufficiently small ϵ we have that,

$$\left|\!\left|\!\left| \sum_{i=1}^{m} w_i \big(g\left(z_i\right) - \widehat{\theta} \big) \big(g\left(z_i\right) - \widehat{\theta} \big)^\top - \operatorname{cov}\left(\theta_g\right) \right|\!\right|\!\right|_{k,op} \geq \frac{\left\| P_k \left(\widetilde{\Delta}\left(w\right) \right) \right\|_2^2}{4\epsilon}.$$

Proof Let $\mathcal{S} = \operatorname{argmax}_{\mathcal{S}' \subset [d], |\mathcal{S}'| \leq k} \left\| \Delta^{\mathcal{S}'} \right\|_2$. Assumptions in the lemma imply that

$$\begin{aligned} \left\| \sum_{i \in \mathcal{B}} w_i \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right) \right\|_2 &= \left\| \widetilde{\Delta}^{\mathcal{S}} - \sum_{i \in \mathcal{G}} w_i \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right) \right\|_2 \\ &= \frac{C_1}{5} \left(L_F + \sqrt{L_{\text{cov}}} \right) \delta - c \left(L_F + \sqrt{L_{\text{cov}}} \right) \delta \\ &= \left(\frac{C_1}{5} - c \right) \left(L_F + \sqrt{L_{\text{cov}}} \right) \delta \end{aligned}$$

where we have used Lemma A.1. Now consider the covariance. We have

$$\sum_{i \in \mathcal{B}} \frac{w_i}{w_b} \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right) \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right) \succcurlyeq \left(\frac{w_i}{w_b} \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right) \right) \left(\frac{w_i}{w_b} \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right) \right)$$

because of the non-negativity of variance. Therefore, because $|\mathcal{B}| \leq 2\epsilon n$, we have

$$\left\| \sum_{i \in \mathcal{B}} \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right) \left(g^{\mathcal{S}}\left(z_i\right) - \theta_g^{\mathcal{S}} \right)^\top \right\|_{\text{op}} \geq \frac{\left\| \widetilde{\Delta}^{\mathcal{S}} \right\|_2^2}{2\epsilon}.$$

Now using our assumption on the covariance, we have

$$\begin{aligned}
&\left\|\sum_{i=1}^{m} w_i \left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)\left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)^{\top}\right\|_{\text{op}} \\
\geq &\left\|\sum_{i\in\mathcal{B}} w_i \left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)\left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)^{\top}\right\|_{\text{op}} - \\
&\left\|\sum_{i\in\mathcal{G}} w_i \left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)\left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)^{\top} - w_g \text{cov}\left(\theta_g^{\mathcal{S}}\right)\right\|_{\text{op}} - \left\|w_b \text{cov}\left(\theta_g^{\mathcal{S}}\right)\right\|_{\text{op}} \\
= &\frac{\left\|\widetilde{\Delta}^{\mathcal{S}}\right\|^2}{2\epsilon} - c\left(L_F^2 + L_{\text{cov}}\right)\delta - 2\epsilon L_{\text{cov}} \\
\geq &\frac{\left\|\widetilde{\Delta}^{\mathcal{S}}\right\|^2}{3\epsilon}
\end{aligned}$$

where in the last inequality we have used the assumption that ϵ is sufficiently small. Lastly, we use the expression

$$\begin{aligned}
&\sum_{i=1}^{m} w_i \left(g^{\mathcal{S}}\left(z_i\right)-\widehat{\theta}^{\mathcal{S}}\right)\left(g^{\mathcal{S}}\left(z_i\right)-\widehat{\theta}^{\mathcal{S}}\right)^{\top} - \text{cov}\left(\theta_g^{\mathcal{S}}\right) \\
= &\sum_{i=1}^{m} w_i \left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)\left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)^{\top} - \text{cov}\left(\theta_g^{\mathcal{S}}\right) - \widehat{\Delta}^{\mathcal{S}}\left(\widetilde{\Delta}^{\mathcal{S}}\right)^{\top} - \widetilde{\Delta}^{\mathcal{S}}\left(\widehat{\Delta}^{\mathcal{S}}\right)^{\top} + \widehat{\Delta}^{\mathcal{S}}\left(\widehat{\Delta}^{\mathcal{S}}\right)^{\top}.
\end{aligned}$$

to obtain

$$\begin{aligned}
&\left\|\sum_{i=1}^{m} w_i \left(g^{\mathcal{S}}\left(z_i\right)-\widehat{\theta}^{\mathcal{S}}\right)\left(g^{\mathcal{S}}\left(z_i\right)-\widehat{\theta}^{\mathcal{S}}\right)^{\top} - \text{cov}\left(\theta_g^{\mathcal{S}}\right)\right\|_{\text{op}} \\
\geq &\left\|\sum_{i=1}^{m} w_i \left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)\left(g^{\mathcal{S}}\left(z_i\right)-\theta_g^{\mathcal{S}}\right)^{\top} - \text{cov}\left(\theta_g^{\mathcal{S}}\right)\right\|_{\text{op}} - 24\left(\left\|\widetilde{\Delta}^{\mathcal{S}}\right\|_2^2\right) \\
\geq &\frac{\left\|\widetilde{\Delta}^{\mathcal{S}}\right\|^2}{4\epsilon}.
\end{aligned}$$

■

Lemma A.6 *Using the same notations and assuming the same conditions as Lemma A.2, we have*

$$\left\|\sum_{i=1}^{m} w_i \left(g\left(z_i\right)-\widehat{\theta}\left(w\right)\right)\left(g\left(z_i\right)-\widehat{\theta}\left(w\right)\right)^{\top} - F\left(\widehat{\theta}\left(w\right)\right)\right\|_{k,op} \geq \frac{\left\|P_k\left(\widetilde{\Delta}\left(w\right)\right)\right\|_2^2}{5\epsilon}.$$

Proof With the same notations in the proof of Lemma A.1, we know

$$\left\| \sum_{i=1}^{m} w_i \left(g^{\mathcal{S}}\left(z_i\right) - \widehat{\theta}^{\mathcal{S}} \right) \left(g^{\mathcal{S}}\left(z_i\right) - \widehat{\theta}^{\mathcal{S}} \right)^{\top} - \operatorname{cov}\left(\theta_g^{\mathcal{S}}\right) \right\|_{\text{op}} \geq \frac{\left\| \widetilde{\Delta}^{\mathcal{S}} \right\|^2}{4\epsilon}.$$

By our assumptions on F, we have

$$\begin{aligned} \left\|\left| F\left(\theta\right) - F\left(\widehat{\theta}\right) \right|\right\|_{\text{k,op}} \leq & L_F \left\| \widehat{\Delta} \right\|_2 + C \left\| \widehat{\Delta} \right\|_2^2 \\ \leq & 5 L_F \left\| \widetilde{\Delta}^{\mathcal{S}} \right\|_2 + 5C \left\| \widetilde{\Delta}^{\mathcal{S}} \right\|_2^2 . \end{aligned}$$

Since $\delta = \Omega\left(\epsilon\right)$, ϵ is larger than any absolute constant, applying triangle inequality to previous two inequalities, we obtain the desired result. ■

A.3. Proofs of Convex Relaxation of Sparse PCA

Theorem A.1 (Theorem A.3) *For a fixed w, the optimal value $\lambda^*\left(w\right)$ of Eqn. (20) satisfies*

$$\lambda^*\left(w\right) \geq \left\|\left| \sum_{i=1}^{m} w_i \big(g\left(z_i\right) - \widehat{\theta}\left(w\right) \big) \big(g\left(z_i\right) - \widehat{\theta}(w) \big)^{\top} - F\big(\widehat{\theta}(w)\big) \right|\right\|_{k,op} .$$

Furthermore, the solution $H^*(w)$ satisfies that there is a universal constant C such that for any $w' \in S_{m,\epsilon}$

$$\begin{aligned} & \operatorname{tr}\Big(\Big(\sum_{i=1}^{m} w_i' \big(g\left(z_i\right) - \widehat{\theta}(w) \big) \big(g\left(z_i\right) - \widehat{\theta}\left(w\right) \big) - F\big(\widehat{\theta}\left(w\right)\big) \Big) H^*(w) \Big) \\ \leq & C \left(k \left\| \mathcal{E}(w') \right\|_{\infty} + \left\| \widehat{\Delta}\left(w\right) \right\|_2^2 + \left(L_F + s \left\| \widetilde{\Delta}\left(w'\right) \right\|_{\infty} \right) \left\| \widehat{\Delta}\left(w\right) \right\|_2 \right) . \end{aligned}$$

Proof Because this is a convex relaxation of sparse PCA, the lower bound is naturally satisfied. For the upper bound, again we use the decomposition

$$\begin{aligned} & \sum_{i=1}^{m} w_i' \left(g\left(z_i\right) - \widehat{\theta} \right) \left(g\left(z_i\right) - \widehat{\theta} \right) - F\left(\widehat{\theta}\right) \\ = & \mathcal{E}\left(w'\right) - \widetilde{\Delta}\left(w'\right) \widehat{\Delta}\left(w\right)^{\top} - \widehat{\Delta}\left(w\right) \widetilde{\Delta}\left(w'\right)^{\top} + \widehat{\Delta}\left(w\right) \widehat{\Delta}\left(w\right)^{\top} + \operatorname{cov}\left(g\right) - F\left(\widehat{\theta}\left(w\right)\right) . \end{aligned}$$

First applying Hölder inequality on trace we have

$$\operatorname{tr}\left(\mathcal{E}\left(w'\right) H^*\left(w\right) \right) \leq \left\| \mathcal{E}\left(w'\right) \right\|_{\infty,\infty} \left\| H^*\left(w\right) \right\|_{1,1} \leq k \left\| \mathcal{E}\left(w'\right) \right\|_{\infty,\infty} .$$

Similarly, we have

$$\begin{aligned} & \operatorname{tr}\left(\left(\widetilde{\Delta}\left(w'\right) \widehat{\Delta}\left(w\right)^{\top} + \widehat{\Delta}\left(w\right) \widetilde{\Delta}\left(w'\right)^{\top} \right) H^*\left(w'\right) \right) \\ \leq & 2k \left\| \widetilde{\Delta}\left(w'\right) \right\|_{\infty} \left\| \widehat{\Delta}\left(w\right) \right\|_2 . \end{aligned}$$

Note $H^*(w)$ belongs to the Fantope $\mathcal{F}^1$ (Overton and Womersley, 1992; Vu et al., 2013), so

$$\operatorname{tr}\left(\widehat{\Delta}(w)\widehat{\Delta}(w)^\top H^*(w)\right) \leq \left\|\widehat{\Delta}(w)\widehat{\Delta}(w)^\top\right\|_{\text{op}} \leq \left\|\widehat{\Delta}(w)\right\|_2^2.$$

Using this property again we have

$$\begin{aligned}\operatorname{tr}\left(\left[\operatorname{cov}(g) - F\left(\widehat{\theta}(w)\right)\right] H^*(w)\right) &\leq \left\|\operatorname{cov}(g) - F\left(\widehat{\theta}(w)\right)\right\|_{\text{op}} \\ &\leq L_F\left\|\widehat{\Delta}(w)\right\|_2 + C\left\|\widehat{\Delta}(w)\right\|_2^2.\end{aligned}$$

Putting these together we obtain the desired result. ■

A.4. Proofs of Ellipsoid Algorithm

We begin with proving the correctness of the separation oracle.

Theorem A.2 (Separation Oracle) *Let w^* denote the weights which are uniform on the uncorrupted points. Suppose Eqn. (15)-Eqn. (16) hold, then there exists a sufficiently large absolute constant C_{good} that if we set $\tau_{sep} = \Omega\left(\left(L_F^2 + L_{\text{cov}}\right)\delta\right)$, Algorithm 2 satisfies*

1. *(*Completeness*) If $w = w^*$, the algorithm outputs "Yes".*
2. *(*Soundness*) If $w \notin C_{C_{good}\left(L_F^2 + L_{\text{cov}}\right)\delta}$, the algorithm outputs a hyperplane $\ell(\cdot)$ such that $\ell(w) \geq 0$. Moreover, if the algorithm ever outputs a hyperplane ℓ, then $\ell(w^*) < 0$.*

Remark: The conditions of this separation oracle is slightly weaker than the traditional ones. However, note that outside $C_{C_{good}\left(L_F^2 + L_{\text{cov}}\right)\delta}$, the separation oracle acts exactly as a separation oracle for w^*.

Proof First, for the completeness, plugging Eqn. (17) and Eqn. (19) into Theorem A.3 and then using Lemma A.1, we directly obtain the desired result. If $w \notin C_{C_{good}\left(L_F^2 + L_{\text{cov}}\right)\delta}$, we can apply the lower bound in Theorem A.3 and use Lemma A.6. See Lemma A.7 for the full proof. When the algorithm outputs a hyperplane, $\ell(w) \geq 0$ follows directly by the optimality of the convex program. Lastly, we use the upper bound of Theorem A.3 to argue $\ell(w^*) < 0$ whenever we outputs a hyperplane (Lemma A.8). ■

Lemma A.7 *If $w \notin \mathcal{C}_{C_{good}\left(L_F^2 + L_{\text{cov}}\right)\delta}$, then $\lambda^* = \Omega\left(\left(L_F^2 + L_{\text{cov}}\right)\delta\right)$.*

Proof Applying the lower bound of Theorem A.3, we have

$$\begin{aligned}&\operatorname{tr}\left(H^*(w)\, G(w)\right) \\ \geq &\left\|\sum_{i=1}^m w_i\left(g(z_i) - \widehat{\theta}(w)\right)\left(g(z_i) - \widehat{\theta}(w)\right)^\top - F\left(\widehat{\theta}\right)\right\|_{\text{k,op}}.\end{aligned}$$

Now if $\left\|\widehat{\Delta}\right\|_2 \geq 5C_1\left(L_F+\sqrt{L_{\text{cov}}}\right)\delta$ where C_1 is defined in Lemma A.2, by Lemma A.6 and Lemma A.1, we have

$$\begin{aligned}&\left\|\sum_{i=1}^{m} w_i\left(g\left(z_i\right)-\widehat{\theta}\left(w\right)\right)\left(g\left(z_i\right)-\widehat{\theta}\left(w\right)\right)^\top - F\left(\widehat{\theta}\right)\right\|_{\text{k,op}}\\ \geq&\frac{\left\|\widehat{\Delta}\right\|_2^2}{5\epsilon}\\ =&\Omega\left(\left(L_F^2+L_{\text{cov}}\right)\delta\right).\end{aligned}$$

On the other hand if $\left\|\widehat{\Delta}\right\|_2 \leq 5C_1\left(L_F+\sqrt{L_{\text{cov}}}\right)\delta$, by Lemma A.1, by definition of $\mathcal{C}_{C_{good}\left(L_F^2+L_{\text{cov}}\right)\delta}$, we have

$$\begin{aligned}&\operatorname{tr}\left(H^*\left(w\right)M\left(w\right)\right)\\ \geq&C_{good}\left(L_F^2+L_{\text{cov}}\right)\delta-\left\|F\left(\theta\right)-F\left(\widehat{\theta}\right)\right\|_{\text{k,op}}\\ \geq&C_{good}\left(L_F^2+L_{\text{cov}}\right)\delta-L_F\left\|\widehat{\Delta}\right\|_2-C\left\|\widehat{\Delta}\right\|_2^2\\ =&\Omega\left(\left(L_F^2+L_{\text{cov}}\right)\delta\right)\end{aligned}$$

where the last step we use the fact that C_{good} is large enough. ■

Lemma A.8 *For any hyperplane ℓ, $\ell\left(w^*\right)<0$.*

Proof We apply the upper bound of Theorem A.3 with $w'=w^*$. Therefore, we only need to upper bound

$$O\left(k\left\|\mathcal{E}\left(w'\right)\right\|_{\infty,\infty}+\left\|\widehat{\Delta}\left(w\right)\right\|_2^2+\left(L_F+s\left\|\widetilde{\Delta}\left(w'\right)\right\|_\infty\right)\left\|\widehat{\Delta}\left(w\right)\right\|_2\right)-\lambda^*\left(w\right)<0.$$

Plugging in our assumptions on w^*, we just need to show

$$C_2\left(\left(L_F^2+L_{\text{cov}}\right)\delta+\left(\left(L_F+\sqrt{L_{\text{cov}}}\right)\delta+L_F\right)\left\|\widehat{\Delta}\left(w\right)\right\|_2+\left\|\widehat{\Delta}\left(w\right)\right\|_2^2\right)-\lambda^*\left(w\right)<0$$

for some absolute constant C_2. If $\left\|\widehat{\Delta}\left(w\right)\right\|_2 \geq 5C_1\left(L_F + \sqrt{L_{\text{cov}}}\right)\delta$ for C_1 defined in Lemma A.2, using the argument in Lemma A.7, we know $\lambda^*\left(w\right) = \Omega\left(\frac{\|\widehat{\Delta}\|_2^2}{\epsilon}\right)$. Therefore, we have

$$\begin{aligned}
\ell\left(w^*\right) &\leq C_2\left(\left(L_F^2 + L_{\text{cov}}\right)\delta + \left(\left(L_F + \sqrt{L_{\text{cov}}}\right)\delta + L_F\right)\left\|\widehat{\Delta}\left(w\right)\right\|_2 + \left\|\widehat{\Delta}\left(w\right)\right\|_2^2\right) - \Omega\left(\frac{\left\|\widehat{\Delta}\left(w\right)\right\|_2^2}{\epsilon}\right) \\
&\leq C_2\left(\left(L_F^2 + L_{\text{cov}}\right)\delta + \left(\left(L_F + \sqrt{L_{\text{cov}}}\right)\delta + L_F\right)\left\|\widehat{\Delta}\left(w\right)\right\|_2\right) - \Omega\left(\frac{\left\|\widehat{\Delta}\left(w\right)\right\|_2^2}{\epsilon}\right) \\
&\leq C_2 L_F\left\|\widehat{\Delta}\left(w\right)\right\|_2 - \Omega\left(\frac{\left\|\widehat{\Delta}\left(w\right)\right\|_2^2}{\epsilon}\right) \\
&< 0
\end{aligned}$$

where the second equality we used $\left\|\widehat{\Delta}\left(w\right)\right\|_2 \geq 5C_1\left(L_F + \sqrt{L_{\text{cov}}}\right)\delta$ and the third we used the fact that $\delta = \Omega\left(\epsilon\right)$. If $\left\|\widehat{\Delta}\left(w\right)\right\|_2 \leq 5C_1\left(L_F + \sqrt{L_{\text{cov}}}\right)\delta$, since $\lambda^*\left(w\right) \geq \tau_{sep} \geq C_3\left(L_F^2 + L_{\text{cov}}\right)\delta$ for C_3 sufficiently large, we have

$$\begin{aligned}
\ell\left(w^*\right) &\leq C_2\left(\left(L_F^2 + L_{\text{cov}}\right)\delta + \left(\left(L_F + \sqrt{L_{\text{cov}}}\right)\delta + L_F\right)\left\|\widehat{\Delta}\left(w\right)\right\|_2 + \left\|\widehat{\Delta}\left(w\right)\right\|_2^2\right) - C_3\left(L_F^2 + L_{\text{cov}}\right)\delta \\
&= -\Omega\left(\left(L_F^2 + L_{\text{cov}}\right)\delta\right) < 0.
\end{aligned}$$

Thus, whenever we output a hyperplane ℓ, $\ell\left(w^*\right) < 0$. ■

Now, by classical convex programming result, after polynomial iterations we can obtain w such that there exists $w' \in C_{C_{good}\left(L_F^2 + L_{\text{cov}}\right)\delta}$, $\left\|w - w'\right\|_\infty \leq \frac{\epsilon\left(\sqrt{L_{\text{cov}}} + L_F\right)}{nD}$. Lemma A.9 shows this w is good enough to make $\widehat{\theta}_g\left(w\right)$ a good estimate. This finishes the proof of Theorem 3.1.

Lemma A.9 *Given w, if there exists $w' \in C_{C_{good}\left(L_F^2 + L_{\text{cov}}\right)\delta}$ such that $\left\|w - w'\right\|_\infty \leq \frac{\epsilon\left(\sqrt{L_{\text{cov}}} + L_F\right)}{mD}$, then $\left\|\widehat{\Delta}\left(w\right)\right\|_2 = O\left(\left(\sqrt{L_{\text{cov}}} + L_F\right)\delta\right)$.*

Algorithm 5 Naive Pruning for Gaussian Mean

```
Input: $\{x_1, \cdots, x_n\}$
For $i, j = 1, \cdots, n$, let $\delta_{ij} = \|x_i - x_j\|_2$.
for $i = 1, \cdots, j$ do
    Let $A_i = \left\{j \in 1, \cdots, n : \delta_{ij} = \Omega\sqrt{d \log\left(n/\tau\right)}\right\}$
    if $|A_i| > 2\epsilon n$ then
        remove $x_i$ from the set.
    end if
end for
```

Proof By the assumptions, we have

$$\begin{aligned}\left\|\widehat{\Delta}\left(w\right)\right\|_2 \leq & 5\left\|P_k\left(\widetilde{\Delta}_S\left(w\right)\right)\right\|_2 \\ = & 5\left\|P_k\left(\widetilde{\Delta}\left(w'\right) + \sum_{i=1}^{m}\left(w_i - w_i'\right)\left(g\left(z_i\right) - \theta_g\right)\right)\right\|_2 \\ \leq & 5\left\|P_k\left(\Delta\left(w'\right)\right)\right\|_2 + \sum_{i=1}^{m}\left|w_i - w_i'\right|\left\|g\left(z_i\right) - \theta_g\right\|_2 \\ = & O\left(\left(\sqrt{L_{\text{cov}}} + L_F\right)\delta\right) + \left(\sqrt{L_{\text{cov}}} + L_F\right)\epsilon \\ = & O\left(\left(\sqrt{L_{\text{cov}}} + L_F\right)\delta\right).\end{aligned}$$

■

Appendix B. Technical Details of Sparse Mean Estimation

In this section we prove Theorem 3.1. Since $F\left(g\right) = I$, a constant function, we know $L_{\text{cov}} = 1$ and $L_F = 0$. We adopt Algorithm 1 in (Diakonikolas et al., 2016a) to achieve the boundedness condition in Theorem 3.1. The pseudocodes are listed in Algorithm 5 for completeness. Maximal inequality of Gaussian random variables shows with probability $1 - \tau$, this procedure does not remove any example sampled from P.

Now we prove the concentration inequalities in Theorem 3.1. Note when $n = \Omega\left(\frac{k^2 \log(d/\tau)}{\epsilon^2}\right)$ Eqn. (15) - (16) can be proved through classical Bernoulli and Gaussian concentration inequalities. For the remaining two, we use the following lemma.

Lemma B.1 *Fix $0 < \epsilon < 1/2$ and $\tau < 1$. There is a $\delta = O\left(\epsilon\sqrt{\log\left(1/\epsilon\right)}\right)$ such that if $x_1, \cdots, x_n \sim N\left(\mu, I\right)$ and $n = \Omega\left(\frac{k\log d + \log(1/\tau)}{\delta^2}\right)$ then for any $w \in S_{n,\epsilon}$ the followings hold:*

$$\max_{\|v\|_0 \le k, \|v\|_2 \le 1} \left| v^\top \left(\sum_{i=1}^n w_i \left(x_i - \mu\right)\left(x_i - \mu\right)^\top - I \right) v \right| \le \delta \tag{23}$$

$$\max_{\mathcal{S} \subset [d], |\mathcal{S}| \le k} \left\| \sum_{i=1}^n w_i \left(x_i^{\mathcal{S}} - \mu\right) \right\|_2 \le \delta. \tag{24}$$

Proof The proof is similar to Lemma 4.5 of (Diakonikolas et al., 2016a). We prove the concentration result for Eqn. (23), Eqn, (24) follows similarly by replacing the classical concentration inequality of covariance by that of mean. Without loss of generality, we assume $\mu = 0$. For any $J \subset [n]$, $|J| = (1-2\epsilon)\,n$, we let w^J be the vector which is given by $w_i^J = \frac{1}{|J|}$ for $i \in J$ and $w_i^J = 0$ otherwise. By convexity, it suffices to show that

$$\mathbb{P}\left[\forall J \subset [n] : |J| = (1-2\epsilon)\,n \text{ and } \max_{\mathcal{S} \subset [d], |\mathcal{S}| \le k} \left\| \sum_{i=1}^n w_i^J x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - I \right\|_2 \ge \delta \right] \le \tau.$$

We first fix τ', $\mathcal{S} \subset [d]$ with $|S| \le k$ and $J \subset [n]$. Using triangle inequality we have

$$\begin{aligned}&\left\| \sum_{i=1}^n w_i^J x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - I \right\|_{op} \\ \le& \left\| \frac{1}{(1-2\epsilon)\,n} \sum_{i=1}^n x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - \frac{1}{(1-2\epsilon)\,n} I \right\|_{op} + \left\| \frac{1}{(1-2\epsilon)\,n} \sum_{i \notin J} x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - \frac{2\epsilon}{1-2\epsilon} I \right\|_{op}.\end{aligned}$$

The first term is small than $\frac{\delta}{2}$ with probability at least $1 - \frac{\tau'}{2}$ if $n = \Omega\left(\frac{k + \log(1/\tau')}{\delta^2}\right)$. Similarly, the second term is smaller than $\delta/2$ with probability at least $1 - \frac{\tau'}{2}$ if $n = \Omega\left(\frac{\epsilon(k + \log(1/\tau'))}{\delta^2}\right)$. Now by union bound over all subset $\mathcal{S} \subset [d]$ with $|\mathcal{S}| \le k$ we have if $n = \Omega\left(\frac{k\log d + \log(1/\tau)}{\delta^2}\right)$

$$\left\| \frac{1}{(1-2\epsilon)\,n} \sum_{i=1}^n x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - \frac{1}{(1-2\epsilon)\,n} I \right\|_{op} \le \frac{\delta}{2}.$$

Similarly, if $n = \Omega\left(\frac{\epsilon(k\log d + \log(1/\tau'))}{\delta^2}\right)$ we have

$$\left\| \frac{1}{(1-2\epsilon)\,n} \sum_{i \notin J} x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - \frac{2\epsilon}{1-2\epsilon} I \right\|_{op} \le \frac{\delta}{2}.$$

Now choosing $\tau' = \begin{pmatrix} n \\ (1-2\epsilon)\,n \end{pmatrix}^{-1} \tau$ and taking union bounds over all J, by our choice of δ and n in the theorem we have

$$\left\| \frac{1}{(1-2\epsilon)\,n} \sum_{i \notin J} x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - \frac{2\epsilon}{1-2\epsilon} I \right\|_{op} \le \frac{\delta}{2}.$$

Algorithm 6 Pruning for Sparse Covariance Estimation

```
Input: $\{x_1, \cdots, x_n\}$
for $i = 1, \cdots, n$ do
    if $\|x_i\|_2 = \Omega\left(d\sqrt{\log(n/\tau)}\right)$. then
        remove $x_i$ from the set.
    end if
end for
```

Our proof is complete. ■

We accompany our upper bound with the following minimax lower bound.

Theorem B.1 (Lower Bound of Sparse Gaussian Mean Estimation) *There are some constants C, c such that*

$$\inf_{\widehat{\mu}} \sup_{x\sim N(\mu,I),\|\mu\|_0\leq s} \sup_{Q} \mathbb{P}\left[\|\widehat{\mu}-\mu\|_2^2 \geq C\left(\frac{s\log(d)}{n} \vee \epsilon^2\right)\right] \geq c.$$

Proof First, the minimax lower bound for no adversary is $\asymp \frac{k\log d}{n}$. Further we know there exist μ_1 and μ_2 with $\|\mu_1\|_0, \|\mu_2\|_0 \leq k$ and $\mathrm{TV}\left(N(\mu_1, I), N(\mu_2, I)\right) \leq \frac{2\epsilon}{1-2\epsilon}$ such that $\|\mu_1-\mu_2\|_2^2 \geq C'\epsilon$ (just consider two vectors each has only one non-zero entry). Now apply Theorem 4.1 of (Chen et al., 2015). ■

Appendix C. Technical Details of Sparse Covariance Estimation

In this section we prove Theorem 3.2. By Theorem 4.15 of (Diakonikolas et al., 2016a) we have the following formula for the $\mathrm{cov}(\Omega)$

$$F(\Omega) = \Omega \otimes \Omega + \mathrm{vec}(\Omega) \otimes \mathrm{vec}(\Omega).$$

Now observe that $\mathrm{tr}(\Sigma) = d$ so for $x_1, \cdots, x_n \sim \mathcal{N}(0, \Sigma)$, using maximal inequality of Gaussian random variables, we have

$$\mathbb{P}\left[\max_i \|x_i\|_2 \geq \Omega\left(d\sqrt{\log(N/\tau)}\right)\right] \leq \tau.$$

Therefore we can apply Algorithm 6 to achieve the boundedness assumption in Theorem 3.1. Lastly, for the concentration bounds, note that Eqn. (15), Eqn. (16) and Eqn. (18) can be proved by polynomial of Gaussian random variables and Eqn. (17) and Eqn. (19) are simple corollaries of Theorem 4.17 of Diakonikolas et al. (2016a) with a union bound over subsets of $[d]$ with cardinality k.

Appendix D. Technical Details of Sparse Linear Regression

In this section we study the sparse linear regression problem. We begin by investigating the basic properties of our model.

Theorem D.1 *If*

$$x \sim N\left(0, I\right), y = x\beta + \xi \textit{ where } \xi \sim N\left(0, 1\right),$$

then we have

$$\begin{aligned} \mathbb{E}\left[yx\right] &= \beta \\ \operatorname{cov}\left[yx\right] &= \left(\|\beta\|_2^2 + 1\right) I + \beta\beta^\top \end{aligned}$$

Proof We first look at the expectation.

$$\begin{aligned} \mathbb{E}\left[yx\right] =& \mathbb{E}\left[x\left(\beta^\top x + \xi\right)\right] \\ &= \mathbb{E}\left[xx^\top\right]\beta + \mathbb{E}\left[x\right]\mathbb{E}\left[\xi\right] \\ &= \beta. \end{aligned}$$

For the covariance note

$$\operatorname{cov}\left[yx\right] = \mathbb{E}\left[y^2 xx^\top\right] - \beta\beta^\top. \tag{25}$$

We expand the first term.

$$\begin{aligned} \mathbb{E}\left[y^2 xx^\top\right] =& \mathbb{E}\left[x\left(x^\top\beta + \xi\right)\left(\beta^\top x + \xi\right)x^\top\right] \\ =& \mathbb{E}\left[xx^\top\beta\beta^\top xx^\top\right] + \mathbb{E}\left[\xi^2 xx^\top\right] \\ =& \mathbb{E}\left[xx^\top\beta\beta^\top xx^\top\right] + I. \end{aligned}$$

where we have used the independence of x and ξ to cancel out the cross terms. Now consider a single coordinate of $\mathbb{E}\left[xx^\top\beta\beta^\top xx^\top\right]$, using Isserlis's theorem we have

$$\begin{aligned} \mathbb{E}\left[e_i^\top xx^\top\beta\beta^\top xx^\top e_j\right] &= 2\mathbb{E}\left[e_i^\top xx^\top\beta\right]\mathbb{E}\left[\beta^\top xx^\top e_j\right] + \mathbb{E}\left[e_i^\top xx^\top e_j\right]\mathbb{E}\left[\beta^\top xx^\top\beta\right] \\ &= \begin{cases} 2\beta_i^2 + \|\beta\|_2^2 & \text{if } i = j \\ 2\beta_i\beta_j & \text{if } i \neq j. \end{cases} \end{aligned}$$

Note this implies

$$\mathbb{E}\left[xx^\top\beta\beta^\top xx^\top\right] = \|\beta\|_2^2 I + 2\beta\beta^\top.$$

Therefore, we have

$$\operatorname{cov}\left[yx\right] = \left(\|\beta\|_2^2 + 1\right) I + \beta\beta^\top.$$

■

With these expressions at hand, it is easy to upper bound L_F and L_{cov}.

Corollary D.1 *Under the same assumptions as Theorem D.1, we have*

$$\left\|\operatorname{cov}\left(yx\right)\right\|_{op} \leq 2\left\|\beta\right\|_2^2 + 1.$$

Further, if we define $F\left(\widehat{\beta}\right) = \left(\|\beta\|_2^2 + 1\right) I + \beta\beta^\top$, *then it satisfies*

$$\left\|F\left(\beta\right) - F\left(\widehat{\beta}\right)\right\|_{op} \leq 4\left\|\beta\right\|_2\left\|\beta - \widehat{\beta}\right\|_2 + 2\left\|\beta - \widehat{\beta}\right\|_2^2. \tag{26}$$

Algorithm 7 Pruning for Sparse Linear Regression

```
Input: ${(y_1, x_1), \cdots, (y_n, x_n)}$
for $i = 1, \cdots, n$ do
    if $\|x_i\|_2 = \Omega\left(d\sqrt{\log(n/\tau)}\right)$ or $|y_i| = \Omega\left(\left(\rho^2+1\right)\sqrt{\log(n/\tau)}\right)$ then
        remove $(y_i, x_i)$ from the set.
    end if
end for
```

Proof For the operator norm, of the covariance, using triangle inequality, we have

$$\begin{aligned}\|\mathrm{cov}(yx)\|_{op} &= \left\|\left(\|\beta\|_2^2+1\right)I+\beta\beta^\top\right\|_{op}\\ &\leq 2\|\beta\|_2^2+1.\end{aligned}$$

Now for F, note we can express it as sum of terms involves difference of β and $\widehat{\beta}$.

$$\begin{aligned}&F(\beta)-F\left(\widehat{\beta}\right)\\ =&2\beta^\top\left(\beta-\widehat{\beta}\right)I+\beta\left(\beta-\widehat{\beta}\right)^\top+\left(\beta-\widehat{\beta}\right)\beta^\top-\left\|\beta-\widehat{\beta}\right\|_2^2 I-\left(\beta-\widehat{\beta}\right)\left(\beta-\widehat{\beta}\right)^\top.\end{aligned}$$

Therefore, using triangle inequality on the operator norm, we have

$$\left\|F(\beta)-F\left(\widehat{\beta}\right)\right\|_{op}\leq 4\|\beta\|_2\left\|\beta-\widehat{\beta}\right\|_2+2\left\|\beta-\widehat{\beta}\right\|_2^2.$$

■

Now to obtain the boundedness assumption, we can use the procedure in Algorithm 7. Again, maximal inequality of Gaussian random variables shows with probability $1-\tau$, this procedure does not remove any example sampled from P.

It remains to prove the concentration bounds. When $n = \Omega\left(\frac{k^2\log(d/\tau)}{\epsilon^2}\right)$, Eqn. (15), Eqn. (16) and Eqn. (18) can be proved through classical Bernoulli and Gaussian concentration inequalities. For the remaining two, the following lemma suffices.

Lemma D.1 (Concentration bounds for Sparse Linear Regression) *Suppose for $i = 1, \cdots, n$, let*

$$x_i \sim N(0, I), y_i = x_i\beta + \xi_i \text{ where } \xi_i \sim N(0,1).$$

Then if $n = \Omega\left(\frac{k\log(d/\tau)}{\epsilon^2}\right)$, then there is a $\delta = O\left(\epsilon\log^2(1/\epsilon)\right)$ that with probability at least $1-\tau$, we have for any subset $\mathcal{S}\subset[d]$, $|\mathcal{S}|\leq k$ and any $w\in S_{n,\epsilon}$, the followings hold

$$\left\|\sum_{i=1}^n w_i y_i x_i^{\mathcal{S}}-\beta^{\mathcal{S}}\right\|_2 \leq \delta\left(\|\beta\|_2+1\right)$$

$$\left\|\sum_{i=1}^n w_i\left(y_i x_i^{\mathcal{S}}-\beta^{\mathcal{S}}\right)\left(y_i x_i^{\mathcal{S}}-\beta^{\mathcal{S}}\right)^\top-\left(1+\|\beta\|_2^2\right)I_s-\beta^{\mathcal{S}}\left(\beta^{\mathcal{S}}\right)^\top\right\|_{op}\leq\delta\left(\|\beta\|_2^2+1\right).$$

Proof We will prove the covariance the concentration for the covariance. The mean is very similar. note that

$$\begin{aligned}
&\sum_{i=1}^{n} w_i \left(y_i x_i^{\mathcal{S}} - \beta^{\mathcal{S}}\right)\left(y_i x_i^{\mathcal{S}} - \beta^{\mathcal{S}}\right)^\top - \left(1 + \|\beta\|_2^2\right) I_s - \beta_{\mathcal{S}}\left(\beta_{\mathcal{S}}\right)^\top \\
=&\sum_{i=1}^{n} w_i x_i^{\mathcal{S}} \left(x_i\right)^\top \beta\beta^\top x_i \left(x_i^{\mathcal{S}}\right)^\top - \left(\|\beta\|_2^2 I + 2\beta^{\mathcal{S}}\left(\beta^{\mathcal{S}}\right)^\top\right) && (27)\\
&+2\sum_{i=1}^{n} w_i \xi_i x_i^{\mathcal{S}} \beta^\top x_i \left(x_i^{\mathcal{S}}\right)^\top && (28)\\
&+2\beta^{\mathcal{S}}(\beta^{\mathcal{S}})^\top - \sum_{i=1}^{n} w_i \beta^\top x_i x_i^{\mathcal{S}} (\beta^{\mathcal{S}})^\top && (29)\\
&+\sum_{i=1}^{n} w_i \xi_i^2 x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - I && (30)
\end{aligned}$$

We prove the concentration of Eqn. (27), the Eqn. (28) and Eqn. (30) can be proved using similar arguments. Since β is k-sparse, it is sufficient to prove that for any $\mathcal{S}' \subset [d]$, $|\mathcal{S}'| \le k$, $\mathcal{S} = \mathcal{S}' \cup \mathcal{S}^*$ where $\mathcal{S}^*$ is the support of β, the following holds

$$\left\|\sum_{i=1}^{n} w_i x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top \beta^{\mathcal{S}} \left(\beta^{\mathcal{S}}\right)^\top x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - \left(\|\beta\|_2^2 I + 2\beta^{\mathcal{S}}\left(\beta^{\mathcal{S}}\right)^\top\right)\right\|_{op} \le \delta \|\beta\|_2^2 .$$

Now fix $v \in \mathbb{R}^{|S|}$ with $\|v\|_2 = 1$. Define the polynomial $p_v(x) = v^\top x^{\mathcal{S}} \left(x^{\mathcal{S}}\right)^\top \beta^{\mathcal{S}}$. By the same argument in the proof of Theorem 4.17 of Diakonikolas et al. (2016a), under our assumption on δ if $n = \Omega\left(\frac{\log(1/\tau')}{\epsilon^2}\right)$, for any $w \in S_{n,\epsilon}$, with probability $1 - \tau'$

$$\left|\sum_{i=1}^{n} w_i p_v^2(x_i) - \mathbb{E}\left[p_v^2(x)\right]\right| \le \delta \|\beta\|_2^2 .$$

Now take union bound over $\frac{1}{3}$-net of the surface of unit ball of dimension $|S|$ and then take union bound over $\mathcal{S} \in [d]$, we obtain our desired result. Note when $\|\beta\|_2^2 \ge 1$, the error in Eqn. (27) will dominate the other two. On the other hand, if $\|\beta\|_2^2 \le 1$, Eqn. (30) will dominate. Therefore our bound has a $\left(\|\beta\|_2^2 + 1\right)$ factor. ■

Appendix E. Technical Details of Generalized Linear Models

In this section we consider the generalized linear model (GLM). Our derivation heavily depends on the following seminal result from Stein.

Theorem E.1 (Stein's identity (Stein, 1971)) *Let $x \sim N(0, I)$ and G a function satisfying some regularity conditions, then*

$$\mathbb{E}\left[G(x) \cdot x\right] = \mathbb{E}\left[\nabla_x G(x)\right].$$

Algorithm 8 Pruning for Generalized Linear Models

```
Input: $\{(y_1, x_1), \cdots, (y_n, x_n)\}$
for $i = 1, \cdots, n$ do
    if $\|x_i\|_2 = \Omega\left(d\sqrt{\log(n/\tau)}\right)$ or $|y_i| = \Omega\left(u(0) + (\rho^2+1)\sqrt{\log(n/\tau)}\right)$ then
        Remove $(y_i, x_i)$ from the set.
    end if
end for
```

We first investigate the basics properties of GLM.

Theorem E.2 *If $x \sim N(0, I)$ and $y = u(x\beta) + \xi$ where $\xi \sim \mathcal{N}(0, I)$, then we have*

$$\mathbb{E}[yx] = \mathbb{E}\left[\triangledown_{x'} u(x')\right] \cdot \beta,$$

$$\mathbb{E}\left[(yx - \mathbb{E}[yx])(yx - \mathbb{E}[yx])^\top\right] = \mathbb{E}\left[(1 + u^2(x'))I + \left(2u(x')\triangledown^2_{x'} u(x') + \triangledown_{x'} u(x')^2\right)\beta\beta^\top\right].$$

where $x' = x\beta$.

Proof For the first moment, choose $G(x) = u(x\beta)$ we directly have the result. For the covariance, note it is suffice to prove the second moment:

$$\mathbb{E}\left[(yx)(yx)^\top\right] = \mathbb{E}\left[(1 + u^2(x'))I + 2\left(u(x')\triangledown^2_{x'} u(x') + \triangledown_{x'} u(x')^2\right)\beta\beta^\top\right].$$

Write $y = u(x') + \xi$, since $\mathbb{E}\left[\xi^2 xx^\top\right] = I$, we just need to prove

$$\mathbb{E}\left[(u(x')x)(u(x')x)^\top\right] = \mathbb{E}\left[u^2(x')I + 2\left(u(x')\triangledown^2_{x'} u(x') + \triangledown_{x'} u(x')^2\right)\beta\beta^\top\right].$$

Choose $G(x) = u^2(x') \cdot x$ in Stein's identity, we have

$$\mathbb{E}\left[(u(x')x)(u(x')x)^\top\right] = \mathbb{E}\left[g^2(x')\right]I + 2\mathbb{E}\left[u(x')\triangledown_{x'} u(x') \cdot x\right]\beta^\top.$$

Not surprisingly, we can define $G(x) = u(x')\triangledown_{x'} u(x')x$ and apply Stein's identity again to obtain the desired result. ■

Remark: The expression for linear regression can be derived similarly using Stein's identity.

By this expression, we can define

$$F(\beta) = \left(\frac{1 + \mathbb{E}[u^2(x')]}{(\mathbb{E}[\nabla_{x'} u(x')])^2}\right)I + \left(\frac{\mathbb{E}[2u(x')\nabla^2_x u(x') + (\nabla u(x'))^2]}{(\mathbb{E}[\nabla'_x u(x')])^2}\right)\beta\beta^T.$$

as the formula for the covariance. This expression implies that it has the same L_F and L_{cov} as linear regression up to constant factors.

By maximal inequality of Gaussian and Lipschitz condition of u, it is easy to show Algorithm 8 will not remove any sample from P with probability at least $1 - \tau$. now it remains to prove concentrations for yx and $y^2 xx^\top$. When $n = \Omega\left(\frac{k^2\log(d/\tau)}{\epsilon^2}\right)$ Eqn. (15), Eqn. (16) and Eqn. (18) can be proved through classical Bernoulli concentration inequality and Lipschitz function of Gaussian variable concentration inequality. Now we prove the remaining two concentration inequalities. The technique we used is very similar to the proof of Theorem 4.17 of (Diakonikolas et al., 2016a).

Lemma E.1 *Suppose for $i = 1, \cdots, n$,*

$$x_i \sim N(0, I), y_i = u(x_i\beta) + \xi_i \text{ where } \xi_i \sim N(0, 1).$$

with $\|\beta\|_0 \leq s$ and u is a known link function with $u(0) = O(1)$ and 1-Lipschitz. If $n = \Omega\left(\frac{k\log(d/\tau)}{\epsilon^2}\right)$, then there is a $\delta = O\left(\epsilon \log^2\left(\frac{1}{\epsilon}\right)\right)$ that with probability at least $1 - \tau$ we have for any subset $\mathcal{S} \subset [d]$, $|\mathcal{S}| \leq k$ and for any $w \in S_{n,\epsilon}$, we have

$$\left\|\sum_{i=1}^{n} w_i y_i x_i^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right\|_2 = \delta\left(\|\beta\|_2 + 1\right),$$

$$\left\|\sum_{i=1}^{n} w_i \left(y_i x_i^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)\left(y_i x_i^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)^\top - \mathbb{E}\left[\left(yx^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)\left(yx^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)^\top\right]\right\|_{op} \leq \delta\left(\|\beta\|_2^2 + 1\right)$$

where $x' = x\beta$.

Proof We will prove the covariance concentration because the mean concentration is quite similar. Similar to Theorem 3.3, we can divide the expression into 5 parts

$$\begin{aligned}
&\sum_{i=1}^{n} w_i \left(y_i x_i^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)\left(y_i x^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)^\top - \mathbb{E}\left[\left(yx^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)\left(yx^{\mathcal{S}} - \mathbb{E}\left[yx^{\mathcal{S}}\right]\right)^\top\right] \\
=&\sum_{i=1}^{n} w_i \left(u(x_i\beta) - u(0)\right)^2 x^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - \mathbb{E}\left[\left(u(x\beta) - u(0)\right)^2 x^{\mathcal{S}} \left(x^{\mathcal{S}}\right)^\top\right] \qquad (31)\\
+&2\sum_{i=1}^{n} w_i u(0) u(x_i\beta) x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - 2u(0)\mathbb{E}\left[u(x\beta) x^{\mathcal{S}} \left(x^{\mathcal{S}}\right)^\top\right] \\
+&\sum_{i=1}^{n} w_i u^2(0) x^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top - u^2(0) I \\
+&2\sum_{i=1}^{n} w_i \xi_i u(x\beta) x_i^{\mathcal{S}} \left(x_i^{\mathcal{S}}\right)^\top \\
+&\sum_{i=1}^{n} w_i \xi_i^2 x^{\mathcal{S}} \left(x^{\mathcal{S}}\right)^\top - I.
\end{aligned}$$

Again we will prove the concentration for Eqn. (31), the remaining terms can be bounded similarly. Now fix $\mathcal{S}' \subset [d]$ and let $\mathcal{S} = \mathcal{S}' \cup S^\star$ where $S^\star$ is the support of β. For a fixed $v \in \mathbb{R}^{|S|}$, $\|v\|_2 = 1$, define

$$p_v(x) = \left(u\left(x^{\mathcal{S}}\beta\right) - u(0)\right)\left(v^\top x^{\mathcal{S}}\right). \qquad (32)$$

For some fixed large enough constant c, by basic Gaussian concentration inequalities we have

$$\mathbb{P}\left[\left(v^\top x^{\mathcal{S}}\right)^2 \geq \sqrt{c}\log\left(\frac{1}{\epsilon}\right)\right] = O(\epsilon).$$

Similarly, using Lipschitz condition we have

$$\begin{aligned}&\mathbb{P}\left[\left(u\left(x^{\mathcal{S}}\beta\right)-u\left(0\right)\right)^2\geq\sqrt{c}\log\frac{1}{\epsilon}\right]\\ \leq&\mathbb{P}\left[\left|x\beta\right|^2\geq\sqrt{c}\log\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right]\\ =&O\left(\epsilon\right)\end{aligned}$$

Therefore, we have

$$\mathbb{P}\left[p_v^2\left(x\right)\geq c\log^2\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right]=O\left(\epsilon\right).$$

Now applying Hoeffding inequality we have if $n=\Omega\left(\frac{\log(1/\tau)}{\epsilon^2}\right)$, with probability $1-\tau$:

$$\frac{1}{n}\left|\left\{i:p_v^2\left(x_i\right)\geq c\log^2\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right\}\right|\leq 2\epsilon.$$

Now define a distribution $\mathcal{D}$ that for $A_i\sim\mathcal{D}$, $A_i=p_v^2\left(x_i\right)$ if $p^2\left(x_i\right)\leq c\left\|\beta\right\|_2^2\log^2\frac{1}{\epsilon}$ and 0 otherwise. Let α' be the expectation of mean of D. By Hoeffding inequality we can show if $n\geq\Omega\left(\frac{\log(1/\tau)}{\epsilon^2}\right)$ we have with probability $1-\tau$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}A_i-\alpha'\right|=O\left(\epsilon\right).$$

Now let $\alpha=\mathbb{E}\left[p_v^2\left(x\right)\right]=O\left(\left\|\beta\right\|_2^2\right)$. We have

$$\begin{aligned}\left|\alpha'-\alpha\right|&=\mathbb{E}_{x\sim N(0,I)}\left[p_v^2\left(x\right)1_{p_v^2(x)\geq\left(c\log^2\frac{1}{\epsilon}\right)}\right]\\&=\int_{\sqrt{c}\log\frac{1}{\epsilon}}^{\infty}t^2\mathbb{P}\left[p_v^2\left(x\right)\geq t^2\right]dt\\&=\int_{\sqrt{c}\log\frac{1}{\epsilon}}^{\infty}t^2\mathbb{P}\left[\left(x^{\mathcal{S}}\beta\right)^2\left(x^{\mathcal{S}}v\right)^2\geq t^2\right]dt\\&=O\left(c\log^2\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right).\end{aligned}$$

Therefore we have if $n=\Omega\left(\frac{\log\frac{1}{\tau}}{\epsilon^2}\right)$, with probability $1-\tau$

$$\left|\frac{1}{n}\sum_{i=1}^{n}A_i-\alpha\right|=O\left(\epsilon\log^2\frac{1}{\epsilon}\right).$$

Now condition on the followings :

$$\frac{1}{n}\left|\left\{i:p_v^2\left(x_i\right)\geq c\log^2\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right\}\right|\leq 2\epsilon$$

$$\left|\frac{1}{n}\sum_{i=1}^{n}A_i-\alpha\right|=O\left(\epsilon\log^2\frac{1}{\epsilon}\right).$$

Define $J^\star$ the largest $2\epsilon n$ indices of $p_v^2\left(x_i\right)$s and $J_1^\star = \left\{i : p_v^2\left(x_i\right) \geq c\log^2\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right\}$. By the conditions, we known $J_1^\star \subset J^\star$ and

$$\left|\frac{1}{n}\sum_{i\notin J_1^\star} p_v^2\left(x_i\right) - \alpha\right| = O\left(\epsilon\log^2\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right).$$

For any index set I with $|I| = (1-2\epsilon)\,n$, divide $[n]\backslash I = J^+ \cup J^-$ where $J^+ = \left\{i\in I : p_v^2\left(x_i\right) \geq \alpha\right\}$ and $J^- = \left\{i \in I : p_v^2\left(x_i\right) < \alpha\right\}$. First we prove the upper bound

$$\begin{aligned}
&\frac{1}{(1-2\epsilon)\,n}\sum_{i\in I} w_i\left(p^2\left(x_i\right)-\alpha\right)\\
\leq&\frac{1}{(1-2\epsilon)\,n}\sum_{i\in I\cup J^+}\left(p_v^2\left(x_i\right)-\alpha\right) - \frac{1}{(1-2\epsilon)\,n}\sum_{i\in J^-}\left(p_v^2\left(x_i\right)-\alpha\right)\\
\leq&\frac{1}{(1-2\epsilon)\,n}\left|\sum_{i=1}^{n} p_v^2\left(x_i\right)-\alpha\right| + \frac{2}{(1-2\epsilon)\,n}\left|\sum_{i\in J^-}\left(p_v^2\left(x_i\right)-\alpha\right)\right|\\
=&O\left(\epsilon\left\|\beta\right\|_2^2\right) + \frac{\left|J^-\right|}{(1-2\epsilon)\,n}\alpha\\
=&O\left(\epsilon\left\|\beta\right\|_2^2\right)
\end{aligned}$$

where in the fourth line we used concentration inequality of Lipschitz function of Gaussian random variables. For the lower bound

$$\begin{aligned}
\frac{1}{(1-2\epsilon)\,n}\sum_{i\in I}\left(p_v^2\left(x_i\right)-\alpha\right) \geq& \frac{1}{(1-2\epsilon)\,n}\sum_{i\notin J_1^\star}\left(p_v^2\left(x_i\right)-\alpha\right)\\
\geq& -O\left(\epsilon\log^2\frac{1}{\epsilon}\left\|\beta\right\|_2^2\right).
\end{aligned}$$

Note this holds for any I and by convexity for any $w\in S_{n,\epsilon}$ we can conclude that Eqn. (31) holds for fixed S and v. Now take union bounds over $\frac{1}{3}$-net of the surface of unit ball of dimension $|S|$ and subsets of $[d]$ with cardinality $2s$, we obtain the desired result.

Similar to sparse linear regression, the final bound depends on whether $\left\|\beta\right\|_2^2$ is larger than 1 or not, which leads to the form of our bound. ■

Appendix F. Technical Details of Logistic-type Models

In this section we consider the generalized linear model for binomial label.

Theorem F.1 *Suppose $x \sim N(0,I)$ and $y = u(x\beta) + \xi(x\beta)$ where g is a known link function and*

$$\xi(x\beta) = \begin{cases} -u(x\beta) & w.p \quad 1-u(x\beta)\\ 1-u(x\beta) & w.p \quad u(x\beta)\end{cases}$$

Algorithm 9 Pruning for Logistic-type Models

```
Input: $\{(y_1, x_1), \cdots, (y_n, x_n)\}$
for $i = 1, \cdots, n$ do
    if $\|y_i x_i\|_2 = \Omega\left(|\mathbb{E}[\triangledown_{x'} u(x')]| \sqrt{d \log(n/\tau)}\right)$. then
        Remove $(y_i, x_i)$ from the set.
    end if
end for
```

then we have

$$\mathbb{E}[yx] = \mathbb{E}\left[\triangledown_{x'} u(x')\right] \cdot \beta$$
$$\mathbb{E}\left[(yx - \mathbb{E}[yx])(yx - \mathbb{E}[yx])^\top\right] = \mathbb{E}\left[u(x')\right] I + \left(\mathbb{E}\left[\triangledown^2_{x'} u(x')\right] - \mathbb{E}\left[\triangledown_{x'} u(x')\right]^2\right) \cdot \beta\beta^\top$$

where $x' = x\beta$.

Proof This is a simple application of Stein's identity. The derivation is similar to sparse generalized linear model. ■

To achieve the boundedness condition, we resort to Algorithm 9. Finally, the concentration bounds can be proved using the exactly same arguments in Sec. E. Notice that the function p_v defined in Eqn. (32) has a better concentration property:

$$\mathbb{P}\left[p_v^2(x) \geq c \log\left(\frac{1}{\epsilon}\right)\right] = O(\epsilon)$$

because of the boundedness of $u(\cdot)$. This fact leads to a slightly stronger bound than that of generalized linear models.

Appendix G. Technical Details for Sparse PCA

G.1. Implementing DETECTROBUSTSPCA

We first show that the algorithm presented above can be efficiently implemented. Indeed, one can show that by taking the dual of the SDP defining the $\|\cdot\|_{\mathcal{X}_k}$ norm, this problem can be re-written as an SDP with (up to constant factor blowups) the same number of constraints and variables, and therefore we may solve it using traditional SDP solver techniques.

Alternatively, one may observe that to optimize Algorithm 4 via ellipsoid or cutting plane methods, it suffices to, given $w \in S_{n,\epsilon}$, produce a separating hyperplane for the constraint (21). This is precisely what dual norm maximization allows us to do efficiently. It is straightforward to show that the volume of $S_{n,\epsilon} \times \mathcal{X}_k$ is at most exponential in the relevant parameters. Therefore, by the classical theory of convex optimization, (see e.g. Grötschel et al. (2012)), for any ξ, we may find a solution w' and γ' so that $\|w' - w^*\|_\infty \leq \xi$ and γ' so that $|\gamma - \gamma'| < \xi$ for some exact minimizer w^*, where γ is the true value of the solution, in time $\mathrm{poly}(d, n, 1/\epsilon, \log 1/\xi)$,

Neither approach will in general give exact solutions, however, both can achieve inverse polynomial accuracy in the parameters in polynomial time. It should be evident from the analysis that these errors will not affect the correctness of our algorithm, and thus we will ignore these issues of numerical precision throughout the remainder of this section, and assume we work with exact γ.

Observe that in general it may be problematic that we don't have exact access to the minimizer w^*, since some of the X_i may be unboundedly large (in particular, if it's corrupted) in norm. However, we only use information about γ. Since γ lives within a bounded range, and our analysis is robust to small changes to γ, these numerical issues do not change anything in the analysis.

G.2. Proof of Theorem 3.2

For simplicity of notation, we will say that $x_1, \dots, x_n$ is an ϵ-corrupted set of samples from $N(0, \Sigma)$ if they are distributed as in (6). We also let $w^g = |\mathcal{G}|/n$ and $w^b = |\mathcal{B}|/n$.

We now show that Algorithm 4 provides the guarantees required for Theorem 3.2. We first show that if we are in Case 1, then γ is small:

Lemma G.1 *Let $\rho, \delta > 0$. Let ϵ, η be as in Theorem 3.2. Let $x_1, \dots, x_n$ be an ϵ-corrupted set of samples from $N(0, I)$ of size n, where n is as in Theorem 3.2. Then, with probability $1 - \delta$, we have $\gamma \leq \rho/2$.*

Proof Let w be the uniform weights over the uncorrupted points. Then it follows from Theorem A.3 that $\left\|\sum_{i \in \mathcal{G}} w_i (X_i X_i^T - I)\right\|_{\mathcal{X}_k}^* \leq O(\eta)$ with probability $1 - \delta$. Since $w \in S_{n,\epsilon}$, this immediately implies that $\gamma \leq O(\rho)$. By setting constants appropriately, we obtain the desired guarantee. ■

We now show that if we are in Case 2, then γ must be large. We first require the following concentration bounds:

Theorem G.1 *Fix $\epsilon, \delta > 0$. Let $x_1, \dots x_n \sim N(0, I)$, where*

$$n = \Omega\left(\frac{\min(d, k^2) + \log \binom{d^2}{k^2} + \log 1/\delta}{\epsilon^2}\right) .$$

Then

$$\left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^T - I\right\|_{\mathcal{X}_k}^* \leq \epsilon .$$

Let us first introduce the following definition.

Definition 1 *A* symmetric sparsity pattern *is a set S of indices $(i, j) \in [d] \times [d]$ so that if $(i, j) \in S$ then $(j, i) \in S$. We say that a symmetric matrix $M \in \mathbb{R}^{d \times d}$ respects a symmetric sparsity pattern S if* $\mathrm{supp}(M) = S$.

With this definition, we now show:

Lemma 2 *Let $n = O\left(\frac{\min(d,k^2) + \log \binom{d^2}{k^2} + \log 1/\delta}{\epsilon^2}\right)$. Then, with probability $1 - \delta$, the following holds:*

$$|\mathrm{tr}((\widehat{\Sigma} - I)X)| \leq O(\epsilon), \text{ for all symmetric } X \text{ with } \|X\|_0 = k^2 \text{ and } \|X\|_F \leq 1. \tag{33}$$

Proof Fix any symmetric sparsity pattern S so that $|S| \leq k^2$. By classical arguments one can show that there is a $(1/3)$-net over all symmetric matrices X with $\|X\|_F = 1$ respecting S of size at most $9^{O(\min(d,k^2))}$. By a basic net argument, we know that for any δ', we know that except with probability $1 - \delta'$, if we take $n = O\left(\frac{\min(d,k^2)+\log 1/\delta'}{\epsilon^2}\right)$ samples, then for all symmetric X respecting S so that $\|X\|_F \leq 1$, we have $|\mathrm{tr}((\widehat{\Sigma} - I)X)| \leq \epsilon$. The claim then follows by further union bounding over all $O\left(\binom{d^2}{k^2}\right)$ symmetric sparsity patterns S with $|S| \leq k^2$. ■

We will also require the following structural lemma.

Lemma 3 *Any PSD matrix X so that $\mathrm{tr}(X) = 1$ and $\|X\|_1 \leq k$ can be written as*

$$X = \sum_{i=1}^{O(n^2/k^2)} Y_i ,$$

where each Y_i is symmetric, have $\sum_{i=1}^{O(n^2/k^2)} \|Y_i\|_F \leq 4$, and each Y_i is k^2-sparse.

Proof Observe that since X is PSD, then $\|X\|_F \leq \mathrm{tr}(X) = 1$. For simplicity of exposition, let us ignore that the Y_i must be symmetric for this proof. We will briefly mention how to in addition ensure that the Y_i are symmetric at the end of the proof. Sort the entries of X in order of decreasing $|X_{ij}|$. Let Y_i be the matrix whose nonzeroes are the $ik^2 + 1$ through $(i+1)k^2$ largest entries of X, in the same positions as they appear in X. Then we clearly have that $\sum Y_i = X_i$, and each Y_i is exactly k^2-sparse.[3] Thus it suffices to show that $\sum \|Y_i\|_F \leq 4$. We have $\|Y_1\|_F \leq \|X\|_F \leq 1$. Additionally, we have $\|Y_{i+1}\|_F \leq \frac{1^T|Y_i|1}{k}$, which follows simply because every nonzero entry of Y_{i+1} is at most the smallest entry of Y_i, and each has exactly k^2 nonzeros (except potentially the last one, but it is not hard to see this cannot affect anything). Thus, in aggregate we have

$$\sum_{i=1}^{O(n^2/k^2)} \|Y_i\|_F \leq 1 + \sum_{i=2}^{O(n^2/k^2)} \|Y_i\|_F \leq 1 + \sum_{i=1}^{O(n^2/k^2)} \frac{1^T|Y_i|1}{k} = 1 + \frac{1^T|X|1}{k} \leq 2 ,$$

which is stronger than claimed.

However, as written it is not clear that the Y_i's must be symmetric, and indeed they do not have to be. The only real condition we needed was that the Y_i's (1) had disjoint support, (2) summed to X, (3) are each $\Theta(k^2)$ sparse (except potentially the last one), and (4) the largest entry of Y_{i+1} is bounded by the smallest entry of Y_i. It should be clear that this can be done while respecting symmetry by doubling the number of Y_i, which also at most doubles the bound in the sum of the Frobenius norms. We omit the details for simplicity. ■

Proof [Proof of Theorem G.1] Let us condition on the event that (33) holds. We claim then that for all $X \in \mathcal{X}$, we must have $|\mathrm{tr}((\widehat{\Sigma} - I)X)| \leq O(\epsilon)$, as claimed. Indeed, by Lemma 3, for all $X \in \mathcal{X}$, we have that

$$X = \sum_{i=1}^{O(d^2/k^2)} Y_i ,$$

[3] Technically the last Y_i may not be k^2 sparse but this is easily dealt with, and we will ignore this case here

where each Y_i is symmetric, have $\sum_{i=1}^{O(d^2/k^2)} \|Y_i\|_F \leq 4$, and each Y_i is k^2-sparse. Thus,

$$\begin{aligned}
|\mathrm{tr}((\widehat{\Sigma} - I)X)| &\leq \sum_{i=1}^{O(d^2/k^2)} \left|\mathrm{tr}((\widehat{\Sigma} - I)Y_i)\right| \\
&= \sum_{i=1}^{O(d^2/k^2)} \|Y_i\|_F \left|\mathrm{tr}\left((\widehat{\Sigma} - I)\frac{Y_i}{\|Y_i\|_F}\right)\right| \\
&\overset{(a)}{\leq} \sum_{i=1}^{O(d^2/k^2)} \|Y_i\|_F \cdot O(\epsilon) \\
&\overset{(b)}{\leq} O(\epsilon) \; ,
\end{aligned}$$

where (a) follows since each $Y_i/\|Y_i\|_F$ satisfies the conditions in (33), and (b) follows from the bound on the sum of the Frobenius norms of the Y_i. ■

We now need concentration over all choices of weights. This follows from Theorem **??** and the same technique as in the proof of Lemma B.1, so we omit the proof.

Theorem G.2 *Fix $\epsilon \leq 1/2$ and $\delta \leq 1$, and fix $k \leq d$. There is a $\eta = O(\epsilon\sqrt{\log 1/\epsilon})$ so that if $x_1, \ldots, x_n \sim N(0, I)$ and $n = \Omega\left(\frac{\min(d,k^2)+\log\binom{d^2}{k^2}+\log 1/\delta}{\eta^2}\right)$, then we have*

$$\Pr\left[\exists w \in S_{n,\epsilon} : \left\|\frac{1}{n}\sum_{i=1}^{n} w_i x_i x_i^T - I\right\|_{\mathcal{X}_k}^* \geq \eta\right] \leq \delta \; .$$

Lemma G.2 *Let $\rho, \delta > 0$. Let ϵ, η, n be as in Theorem 3.2. Let $x_1, \ldots, x_n$ be an ϵ-corrupted set of samples from $N(0, I)$ of size n. Then, with probability $1 - \delta$, we have $\gamma \geq (1 - \epsilon)\rho - (2 + \rho)\eta$. In particular, for ϵ sufficiently small, and $\eta = O(\rho)$, we have that $\gamma > \rho/2$.*

Proof Let $\Sigma = I + \rho vv^T$, and let $Y_i = \Sigma^{-1/2}x_i$, so that if Y_i is uncorrupted, then $Y_i \sim N(0, I)$. Let w^* be the optimal solution to (21). By Theorem G.2, we have that with probability $1-\delta$, we can write $\sum_{i=1}^{n} w_i^* Y_i Y_i^T = w^g(I + N) + B$, where $\|N\|_{\mathcal{X}_k}^* \leq \eta$, and $B = \sum_{i\in\mathcal{B}} w_i^* Y_i Y_i^T$. Therefore, we have $\sum_{i=1}^{n} w^* X_i X_i^T = w^g(\Sigma + \Sigma^{1/2}N\Sigma^{1/2}) + \Sigma^{1/2}B\Sigma^{1/2}$. By definition, we have

$$\begin{aligned}
\left\|\sum_{i=1}^{n} w_i(X_i X_i^T - I)\right\|_{\mathcal{X}_k}^* &\geq \langle w^g(\Sigma + \Sigma^{1/2}N\Sigma^{1/2}) + \Sigma^{1/2}B\Sigma^{1/2} - I, vv^T\rangle \\
&\geq w^g\langle(\Sigma + \Sigma^{1/2}N\Sigma^{1/2}), vv^T\rangle - 1 \\
&= w^g(1+\rho) + w^g v^T\Sigma^{1/2}N\Sigma^{1/2}v - 1 \\
&\geq (1-\epsilon)\rho + (1-\epsilon)v^T\Sigma^{1/2}N\Sigma^{1/2}v - \epsilon \; .
\end{aligned}$$

It thus suffices to show that $|v^T\Sigma^{1/2}N\Sigma^{1/2}v| < (1+\rho)\eta$. Since v is an eigenvector for Σ with eigenvalue $1+\rho$, we have that $\Sigma^{1/2}v = \sqrt{\rho+1}\cdot v$ and thus

$$v^T\Sigma^{1/2}N\Sigma^{1/2}v = (1+\rho)v^T N v = (1+\rho)\langle N, vv^T\rangle \leq (1+\rho)\|N\|_{\mathcal{X}_k}^* \leq (1+\rho)\eta \; .$$

■

Lemmas G.1 and G.2 together imply the correctness of DETECTROBUSTSPCA and Theorem 3.2.

G.3. More concentration bounds

Before we can prove correctness of our algorithm for robust recovery, we require a couple of concentration inequalities for the set $\mathcal{W}_k$.

Lemma G.3 *Fix $\epsilon, \delta > 0$. Let $x_1, \ldots, x_n \sim N(0, I)$, where n is as in Theorem 3.3. Then with probability $1 - \delta$*

$$\left\|\left\| \frac{1}{n} \sum_{i=1}^n x_i x_i^T - I \right\|\right\|_{\mathcal{W}_k}^* \leq O(\epsilon) \; .$$

Proof Let $\widehat{\Sigma}$ denote the empirical covariance. Observe that $\mathcal{W}_k \subseteq \bigcup_{i=0}^\infty 2^{-i} \mathcal{X}_{2^{i+1}k}$. Moreover, for any i, by Theorem G.1, if we take

$$\begin{aligned} n &= \Omega\left(\frac{\min(d, (2^{i+1}k)^2) + \log \binom{d^2}{(2^{i+1}k)^2} + \log 1/\delta}{(2^{-i}\epsilon)^2}\right) \\ &= \Omega\left(\frac{\min(d, k^2) + \log \binom{d^2}{k^2} + 2^{2i} \log 1/\delta}{\epsilon^2}\right) \; , \end{aligned}$$

then $|\langle M, \widehat{\Sigma} \rangle| \leq \epsilon$ for all $M \in 2^{-i} \mathcal{X}_{2^{i+1}k}$ with probability $1 - \delta/2$. In particular, if we take

$$n = \Omega\left(\frac{\min(d, k^2) + \log \binom{d^2}{k^2} + \log 1/\delta}{\epsilon^2}\right)$$

samples, then for any i, we have $|\langle M, \widehat{\Sigma} \rangle| \leq \epsilon$ for all $M \in 2^{-1} \mathcal{X}_{2^{i+1}k}$ with probability at least $1 - \delta^{2^{2i}}/2$. By a union bound over all these events, since $\sum_{i=0}^\infty \delta^{2^{2i}} \leq 2\delta$, we conclude that if we take n to be as above, then $|\langle M, \widehat{\Sigma} \rangle| \leq \epsilon$ for all $M \in \bigcup_{i=0}^\infty 2^{-i} \mathcal{X}_{2^{i+1}k}$ with probability $1 - \delta$. Since $\mathcal{W}_k$ is contained in this set, this implies that $\|\widehat{\Sigma} - \Sigma\|_{\mathcal{W}_k}^* \leq O(\epsilon)$ with probability at least $1 - \delta$, as claimed. ■

By the same techniques as in the proofs of Theorem G.2, we can show the following bound. Because of this, we omit the proof for conciseness.

Corollary G.1 *Fix $\epsilon, \delta > 0$. Let $x_1, \ldots, x_n \sim N(0, I)$ where n is as in Theorem G.2. Then there is an $\eta = O(\epsilon \sqrt{\log 1/\epsilon})$ so that*

$$\Pr\left[\exists w \in S_{n,\epsilon} : \left\|\left\| \sum_{i=1}^n w_i x_i x_i^T - I \right\|\right\|_{\mathcal{W}_k}^* \geq \eta \right] \leq \delta \; .$$

G.4. Proof of Theorem 3.3

In the rest of this section we will condition on the following deterministic event happening:

$$\forall w \in S_{n,\epsilon} : \left\| \sum_{i=1}^{n} w_i x_i x_i^T - I \right\|_{\mathcal{W}_{2k}}^{*} \leq \eta \ , \tag{34}$$

where $\eta = O(\epsilon \log 1/\epsilon)$. By Corollary G.1, this holds if we take

$$n = \Omega\left(\frac{\min(d, k^2) + \log \binom{d^2}{k^2} + \log 1/\delta}{\eta_2^2} \right)$$

samples.

The rest of this section is dedicated to the proof of the following theorem, which immediately implies Theorem 3.3.

Theorem G.3 *Fix ϵ, δ, and let η be as in (34). Assume that (34) holds. Let $\widehat{v}$ be the output of* RECOVERYROBUSTSPCA$(X_1, \ldots, X_n, \epsilon, \delta, \rho)$. *Then $L(\widehat{v}, v) \leq O(\sqrt{(1+\rho)\eta/\rho})$.*

Our proof proceeds in a couple of steps. Let $\Sigma = I + \rho v v^T$ denote the true covariance. We first need the following, technical lemma:

Lemma G.4 *Let $M \in \mathcal{W}_k$. Then $\Sigma^{1/2} M \Sigma^{1/2} \in (1+\rho)\mathcal{W}_k$.*

Proof Clearly, $\Sigma^{1/2} M \Sigma^{1/2} \succeq 0$. Moreover, since $\Sigma^{1/2} = I + (\sqrt{1+\rho} - 1) v v^T$, we have that the maximum value of any element of $\Sigma^{1/2}$ is upper bounded by $\sqrt{1+\rho}$. Thus, we have $\|\Sigma^{1/2} M \Sigma^{1/2}\|_1 \leq (1+\rho)\|M\|_1$. We also have

$$\begin{aligned} \mathrm{tr}(\Sigma^{1/2} M \Sigma^{1/2}) &= \mathrm{tr}(\Sigma M) \\ &= \mathrm{tr}(M) + \rho v^T M v \leq 1 + \rho \ , \end{aligned}$$

since $\|M\| \leq 1$. Thus $\Sigma^{1/2} M \Sigma^{1/2} \in (1+\rho)\mathcal{W}_k$, as claimed. ■

Let w^*, A^* be the output of our algorithm. We first claim that the value of the optimal solution is quite small:

Lemma G.5

$$\left\| \sum_{i=1}^{n} w_i^* (x_i x_i^T - I) - \rho A^* \right\|_{\mathcal{W}_{2k}}^{*} \leq \eta(1+\rho) \ .$$

Proof Indeed, if we let w be the uniform set of weights over the good points, and we let $A = v v^T$, then by (34), we have

$$\sum_{i=1}^{n} w_i x_i x_i^T = \Sigma^{1/2}(I + N)\Sigma^{1/2} \ ,$$

where $\|N\|_{\mathcal{X}_k}^* \leq \eta$, and $\Sigma = I + \rho vv^T$. Thus we have that

$$\begin{aligned}\left\|\left\|\sum_{i=1}^n w_i(x_i x_i^T - I) - \rho vv^T\right\|\right\|_{\mathcal{W}_{2k}}^* &= \left\|\left\|\Sigma^{1/2} N \Sigma^{1/2}\right\|\right\|_{\mathcal{W}_{2k}}^* \\ &= \max_{M \in \mathcal{W}_k} \left|\mathrm{tr}(\Sigma^{1/2} N \Sigma^{1/2} M)\right| \\ &= \max_{M \in \mathcal{W}_k} \left|\mathrm{tr}(N \Sigma^{1/2} M \Sigma^{1/2})\right| \\ &\leq (1+\rho)\|N\|_{\mathcal{W}_{2k}}^* ,\end{aligned}$$

by Lemma G.4. ■

We now show that this implies the following:

Lemma G.6 $v^T A^* v \geq 1 - (2+3\rho)\eta/\rho$.

Proof By (34), we know that we may write $\sum_{i=1}^n w_i(X_i X_i^T - I) = w^g \rho vv^T + B - (1-w^g)I + N$, where $B = \sum_{i \in \mathcal{B}} w_i X_i X_i^T$, and $|||N|||_{\mathcal{W}_k}^* \leq (1+\rho)\eta$. Thus, by Lemma G.5 and the triangle inequality, we have that

$$\begin{aligned}\left|\left|\left|w^g \rho vv^T + B - \rho A\right|\right|\right|_{\mathcal{W}_k}^* &\leq \eta + |||N|||_{\mathcal{W}_k}^* + (1-w^g)\,|||I|||_{\mathcal{W}_k}^* + (1-w^g)\,|||\rho A|||_{\mathcal{W}_k}^* \\ &\leq (1+\rho)\eta + \epsilon + \rho\epsilon \\ &\leq (1+2\rho)\eta + \epsilon .\end{aligned}$$

Now, since $vv^T \in \mathcal{W}_k$, the above implies that

$$|w^g \rho + v^T B v - \rho v^T A^* v| \leq (1+2\rho)\eta + \epsilon ,$$

which by a further triangle inequality implies that

$$|\rho(1 - v^T A^* v) + v^T B v| \leq (1+2\rho)\eta + \epsilon + \epsilon\rho \leq (2+3\rho)\eta .$$

Since $0 \leq v^T A^* v \leq 1$ (since $A \in \mathcal{X}_k$) and B is PSD, this implies that in fact, we have

$$0 \leq \rho(1 - v^T A^* v) \leq (2+3\rho)\eta .$$

Hence $v^T A^* v \geq 1 - (2+3\rho)\eta/\rho$, as claimed. ■

Let $\gamma = (2+3\rho)\eta/\rho$. The lemma implies that the top eigenvalue of A^* is at least $1-\gamma$. Moreover, since $A^* \in \mathcal{X}_k$, as long as $\gamma \leq 1/2$, this implies that the top eigenvector of A^* is unique up to sign. By the constraint that $\eta \leq O(\min(\rho, 1))$, for an appropriate choice of constants, we that $\gamma \leq 1/10$, and so this condition is satisfied. Recall that u is the top eigenvector of A^*. Since $\mathrm{tr}(A^*) = 1$ and A^* is PSD, we may write $A^* = \lambda_1 uu^T + A_1$, where u is the top eigenvector of A^*, $\lambda_1 \geq 1-\gamma$, and $\|A_1\| \leq \gamma$. Thus, by the triangle inequality, this implies that

$$\left|\left|\left|\rho(vv^T - \lambda_1 uu^T) + B\right|\right|\right|_{\mathcal{X}_{2k}}^* \leq O(\rho\gamma)$$

which by a further triangle inequality implies that

$$\left\|\!\left|\rho(vv^T - uu^T) + B\right|\!\right\|^*_{\mathcal{X}_{2k}} \leq O(\rho\gamma) \; . \tag{35}$$

We now show this implies the following intermediate result:

Lemma G.7 $(v^T u)^2 \geq 1 - O(\gamma)$.

Proof By Lemma G.6, we have that $v^T A^* v = \lambda_1 (v^T u)^2 + v^T A_1 v \geq 1 - \gamma$. In particular, this implies that $(v^T u)^2 \geq (1 - 2\gamma)/\lambda_1 \geq 1 - 3\gamma$, since $1 - \gamma \leq \lambda \leq 1$. ■

We now wish to control the spectrum of B. For any subsets $S, T \subseteq [d]$, and for any vector x and any matrix M, let x_S denote x restricted to S and $M_{S,T}$ denote the matrix restricted to the rows in S and the columns in T. Let I be the support of u, and let J be the support of the largest k elements of v.

Lemma G.8 $\|\!|B_{I,I}\|\!|_{op} \leq O(\rho\gamma)$.

Proof Observe that the condition (35) immediately implies that

$$\left\|\!\left|\rho(v_I v_I^T - u_I u_I^T) + B_{I,I}\right|\!\right\|_{op} \leq c\rho\gamma \; , \tag{36}$$

for some c, since any unit vector x supported on I satisfies $xx^T \in \mathcal{X}_{2k}$. Suppose that $\|\!|B_{I,I}\|\!|_{\mathrm{op}} \geq C\gamma$ for some sufficiently large C. Then (36) immediately implies that $\left\|\!\left|\rho(v_I v_I^T - u_I u_I^T)\right|\!\right\|_{\mathrm{op}} \geq (C - c)\rho\gamma$. Since $(v_I v_I^T - u_I u_I^T)$ is clearly rank 2, and satisfies $\mathrm{tr}(v_I v_I^T - u_I u_I^T) = 1 - \|u_I\|_2^2 \geq 0$, this implies that the largest eigenvalue of $v_I v_I^T - u_I u_I^T$ is positive. Let x be the top eigenvector of $v_I v_I^T - u_I u_I^T$. Then, we have $x^T(v_I v_I^T - u_I u_I^T)x + x^T B x = (C - c)\rho\gamma + x^T B x \geq (C - c)\rho\gamma$ by the PSD-ness of B. If $C > c$, this contradicts (36), which proves the theorem. ■

This implies the following corollary:

Corollary G.2 $\|u_I\|_2^2 \geq 1 - O(\gamma)$.

Proof Lemma G.8 and (36) together imply that $\|\!|v_I v_I^T - u_I u_I^T\|\!|_{\mathrm{op}} \leq O(\gamma)$. The desired bound then follows from a reverse triangle inequality. ■

We now show this implies a bound on $B_{J \setminus I, J \setminus I}$:

Lemma G.9 $\left\|\!\left|B_{J \setminus I, J \setminus I}\right|\!\right\|_{op} \leq O(\rho\gamma)$.

Proof Suppose $\left\|\!\left|B_{J \setminus I, J \setminus I}\right|\!\right\|_{\mathrm{op}} \geq C\gamma$ for some sufficiently large C. Since u is zero on $J \setminus I$, (35) implies that

$$\|\rho v_{J \setminus I} v_{J \setminus I}^T + B_{J \setminus I, J \setminus I}\| \leq c\rho\gamma \; ,$$

for some universal c. By a triangle inequality, this implies that $\|v_{J \setminus I}\|_2^2 = \|v_{J \setminus I} v_{J \setminus I}^T\| \geq (C - c)\gamma$. Since v is a unit vector, this implies that $\|v_I\|_2^2 \leq 1 - (C - c)\gamma$, which for a sufficiently large C, contradicts Corollary G.2. ■

We now invoke the following general fact about PSD matrices:

Lemma G.10 *Suppose M is a PSD matrix, written in block form as*

$$M = \begin{pmatrix} C & D \\ D^T & E \end{pmatrix} .$$

Suppose furthermore that $\|C\|_{op} \leq \xi$ and $\|E\|_{op} \leq \xi$. Then $\|M\| \leq O(\xi)$.

Proof It is easy to see that $\|M\|_{\text{op}} \leq O(\max(\|C\|_{\text{op}}, \|D\|_{\text{op}}, \|E\|_{\text{op}}))$. Thus it suffices to bound the largest singular value of D. For any vectors ϕ, ψ with appropriate dimension, we have that

$$(\phi^T \ -\psi^T)\, M \begin{pmatrix} \phi \\ -\psi \end{pmatrix} = \phi^T A\phi - 2\phi^T D\psi + \psi^T C\psi \geq 0 \, ,$$

which immediately implies that the largest singular value of D is at most $(\|A\|_{\text{op}} + \|B\|_{\text{op}})/2$, which implies the claim. ■

Therefore, Lemmas G.8 and G.9 together imply:

Corollary G.3 $\left\|v_{I\cup J}v_{I\cup J}^T - u_{I\cup J}u_{I\cup J}^T\right\|_{op} \leq O(\gamma)$.

Proof Observe (35) immediately implies that $\left\|\rho(v_{I\cup J}v_{I\cup J}^T - u_{I\cup J}u_{I\cup J}^T) + B_{I\cup J, I\cup J}\right\|_{\text{op}} \leq O(\rho\gamma)$, since $|I \cup J| \leq 2k$. Moreover, Lemmas G.8 and G.9 with Lemma G.10 imply that $\|B_{I\cup J, I\cup J}\|_{\text{op}} \leq O(\rho\gamma)$, which immediately implies the statement by a triangle inequality. ■

Finally, we show this implies $\|vv^T - u_J u_J^T\| \leq O(\gamma)$, which is equivalent to the theorem.
Proof [Proof of Theorem G.3] We will in fact show the slightly stronger statement, that $\left\|uu^T - v_J v_J^T\right\|_{\text{F}} \leq O(\gamma)$. Observe that since $uu^T - vv^T$ is rank 2, Corollary G.3 implies that $\left\|v_{I\cup J}v_{I\cup J}^T - u_{I\cup J}u_{I\cup J}^T\right\|_{\text{F}} \leq O(\gamma)$, since for rank two matrices, the spectral and Frobenius norm are off by a constant factor. We have

$$\left\|uu^T - vv^T\right\|_{\text{F}}^2 = \sum_{(i,j)\in I\cap J\times I\cap J} (u_i u_j - v_i v_j)^2 + \sum_{(i,j)\in I\times I\setminus J\times J} (v_i v_j)^2 + \sum_{(i,j)\in J\times J\setminus I\times I} (u_i u_j)^2 \, .$$

We have

$$\sum_{(i,j)\in I\cap J\times I\cap J} (u_i u_j - v_i v_j)^2 + \sum_{(i,j)\in J\times J\setminus I\times I} (u_i u_j)^2 \leq \left\|v_{I\cup J}v_{I\cup J}^T - u_{I\cup J}u_{I\cup J}^T\right\|_{\text{op}}^2 \leq O(\gamma) \, ,$$

by Corollary G.3. Moreover, we have that

$$\begin{aligned}
\sum_{(i,j)\in I\times I\setminus J\times J} (v_i v_j)^2 &\leq 2\left(\sum_{(i,j)\in I\times I\setminus J\times J} (v_i v_j - u_i u_j)^2 + \sum_{(i,j)\in I\times I\setminus J\times J} (u_i u_j)^2\right) \\
&\leq 2\left(\left\|v_{I\cup J}v_{I\cup J}^T - u_{I\cup J}u_{I\cup J}^T\right\|_{\text{op}}^2 + \sum_{(i,j)\in I\times I\setminus J\times J} (u_i u_j)^2\right) \\
&\leq 2\left(\left\|v_{I\cup J}v_{I\cup J}^T - u_{I\cup J}u_{I\cup J}^T\right\|_{\text{op}}^2 + \sum_{(i,j)\in J\times J\setminus I\times I} (u_i u_j)^2\right) \\
&\leq O(\gamma) \, .
\end{aligned}$$

since $J \times J$ contains the k^2 largest entries of uu^T. This completes the proof. ■