Conditional Linear Regression for Heterogeneous Covariances

Brendan Juba **Leda Liang**
Washington University in St. Louis

Abstract

Often machine learning and statistical models will attempt to describe the majority of the data. However, there may be situations where only a fraction of the data can be fit well by a linear regression model. Here, we are interested in a case where such inliers can be identified by a Disjunctive Normal Form (DNF) formula. We give a polynomial time algorithm for the conditional linear regression task, which identifies a DNF condition together with the linear predictor on the corresponding portion of the data. In this work, we improve on previous algorithms by removing a requirement that the covariances of the data satisfying each of the terms of the condition have to all be very similar in spectral norm to the covariance of the overall condition.

1 INTRODUCTION

Linear regression is a technique frequently used in statistical and data analysis. The task for standard linear regression is to fit a linear relationship among variables in a data set. Often, the goal is to find the most parsimonious model that can describe the majority of the data. In this work, we consider the situation where only a small portion of the data can be accurately modeled using linear regression. More generally, in many kinds of real-world data, portions of the data of significant size can be predicted significantly more accurately than by the best linear model for the overall data distribution: Rosenfeld et al. (2015) showed that there are attributes that are significant risk factors for gastrointestinal cancer in certain subpopulations, but not in the overall population. Hainline et al. (2019) demonstrated that a variety of standard (real-world) regression benchmarks have portions that are fit significantly better by a different linear model than the best model for the overall data set; Calderon et al. (2020) presented further, similar findings. We will consider cases where linear regression fits well when the data set is conditioned on a simple condition, which is unknown to us. We study the task of finding such a linear model, together with a formula on the data attributes describing the condition, i.e., the portion of the data for which the linear model is accurate.

This problem was introduced by Juba (2017), who gave an algorithm for conditional sparse linear regression, using the maximum residual as the objective. This was extended to the usual squared-error loss (as well as other ℓ_p losses) by Hainline et al. (2019). Juba (2017) also gave an algorithm for the general (non sparse) case that could only find a small fraction of the largest such condition. All these algorithms find conditions describing subpopulations that are a union of some basic subsets of data, selected by "terms." For example, simple families of terms may be obtained by considering the data for which a small set of categorical attributes take some specific values, or based on whether the value of some real attributes lie in specific quantiles of the distribution. Calderon et al. (2020) gave an algorithm for non sparse linear regression that matches the size of the largest condition, but only under a new assumption, that the covariances of the data satisfying each of the terms of the condition have to all be very similar in spectral norm to the covariance of the overall condition.

Uniform covariances across terms is an extremely restrictive assumption: it means that essentially the only difference between populations selected by the various terms may be in their means. Note that the problem presupposes that there is significant heterogeneity in the conditional covariances across the distribution overall, or else the same linear model would be equally accurate across the various subsets; concretely, the risk factors found by Rosenfeld et al. (2015) are a kind of correlation between a factor and the target variable that exists in the identified subpopulation, but not in

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

the larger population. In a real-world data set, there is no reason to expect that the only relationships that exist would involve the target attribute; relationships between other pairs of attributes may appear when we consider one term or another. For example, intuitively, if the data lies on a curved manifold, then conditioning on some attribute taking values that select one portion of the curve or another would alter such relationships since the tangent space changes, and the covariance matrix in any local region of the manifold only has eigenvectors lying in the tangent space. (Note that we only require a common component in the orthogonal subspace for a linear model to exist.) So, Calderon et al's algorithm can only be guaranteed to find highly homogeneous subsets of a distribution that features significant heterogeneity overall. In their work, Calderon et al. concluded with the question of whether or not this new property was necessary to obtain a computationally efficient algorithm.

1.1 Our Contribution

In this work, we answer the questions posed by Calderon et al. (2020) and Juba (2017), solving the form of the task originally sought: we give a polynomial-time algorithm that identifies a condition that covers as much of the distribution as the optimal condition and a linear model which provides a good fit when conditioned on said condition, even if the terms of the condition feature heterogeneous covariances. The only assumptions on the data we use are bounds on the moments of the data itself (including hypercontractivity) and generalizations of the standard Gaussian noise assumption on the subset of the data described by the unknown condition. Note that in regression, the error can be arbitrarily large with arbitrary probability, so bounds on the moments of the data are necessary to empirically estimate the error.

Our algorithm is inspired by the list-decodable subspace recovery algorithm presented by Bakshi and Kothari (2021). Their work uses the sum of squares method to construct an algorithm which addresses robust subspace recovery. We present an analogous algorithm for conditional linear regression. As in Calderon et al. (2020), this is done by using a collection of subsets, which we will call "terms," in place of individual points. We thus make use of the fact that by drawing many examples per term, the noise in the data selected by a term can be better controlled, leading to more accurate estimates.

We stress that in contrast to the guarantee that Bakshi and Kothari obtain for their problem, we can obtain arbitrary accuracy with an algorithm that runs in fixed polynomial time, with an exponent that does not depend on the desired accuracy; we only require a sufficient (polynomial) number of examples from the target distribution, and our running time has only a low-order polynomial dependence on the size of the data set. (The dependence on the dimension, by contrast, while fixed, is a higher degree polynomial due to our use of the sum-of-squares method; reducing this dependence is a key challenge for future work, see Section 4.) In particular, given only the certifiable hypercontractivity assumption, Bakshi and Kothari observed that they do not obtain useful estimates for linear regression in fixed polynomial time: a $O(1/\mu)$-accurate estimate is not meaningful for unit-norm coefficient vectors when $\mu \leq 1/2$.

Moreover, for the related problem of robust linear regression with a minority of inliers (discussed further below), all known approaches require a "certifiable anticoncentration" assumption that is much more restrictive, and cannot be satisfied by distributions supported on lower-dimensional manifolds, or discrete distributions such as the uniform distribution on the hypercube, for example. Indeed, Karmalkar et al. (2019) show that anticoncentration is necessary for the robust linear regression problem, but our work shows that it is not necessary for conditional linear regression. Furthermore, Diakonikolas et al. (2021) gave lower bounds for the d-dimensional robust linear regression problem in the statistical query model of $d^{O(1/\mu)}$ queries. It was thus unclear whether or not a fixed polynomial dependence, in which $1/\mu$ does not appear in the exponent, was even possible for our problem.

1.2 Related Work

Our work is both technically and conceptually related to "list-decodable" linear regression. Classical work in robust statistics (Huber, 1981; Rousseeuw and Leroy, 1987) considers situations where a minority subset of the data consists of "outliers" that should be ignored. Recent works (Diakonikolas et al., 2019; Prasad et al., 2020, e.g.) have proposed methods to robustly estimate parameters for a wide variety of models. In this classical setting, it did not make sense to consider the possibility that a majority of the data could be "outliers," in part because there would no longer be a unique, dominant solution to consider. But, a recently proposed model of "list-decodable" robust statistics (Charikar et al., 2017) (similarly to classical algorithms such as RANSAC (Fischler and Bolles, 1981) for subspace discovery) overcomes this obstacle by permitting a list of possible estimates or models to be produced, provided that the list is not too long (generally, $O(1/\mu)$ estimators for a μ-fraction of the data) and that an accurate estimate appears somewhere in the list. In particular, algorithms for list-decodable linear regression have been

proposed (Raghavendra and Yau, 2020; Karmalkar et al., 2019) (see also Bakshi and Kothari (2021)). Although we have formulated our problem in such a way that we only produce a single arbitrary model as output, we could have returned a list of models as well (or vice-versa, select a suitable model from such a list). The main distinction is that in this line of work, on the one hand, one does not need to produce a DNF that identifies the inliers, in contrast to our setting. Note that without this formula, we cannot tell when we should use one of the models versus another to make predictions for new data. Of course, on the other hand, in these works one is not promised that such a DNF exists, either, and so the approach used in our analysis cannot be used in these problems.

Another similar line of work to conditional linear regression is *selective regression* (El-Yaniv and Wiener, 2012): here as well, the objective is to identify a fraction of the data that can be fit well. But in contrast to our setting, this work was in a "transductive" learning setting where a linear predictor is first identified, and then a data set is given, and finally a *ranking* of that data is produced. The interpretation is that the top μ-fraction of the ranking comprise the data for which the predictor is expected to be most accurate. In contrast, we jointly produce a linear model and a DNF that identifies which further examples, drawn from the same distribution, will be accurately predicted by the model. In *learning with rejection* or *abstention* (Cortes et al., 2016), on the other hand, a formula that selects a subset of the data is identified, but the problem formulation assigns a penalty to each example that is "rejected" – thus, we have a default loss value that our classifier can take in place of the loss that would be incurred by this prediction, and this overall loss is minimized over the entire data set. The cost of rejection here takes the place of the probability of the subset μ in our problem.

All of these works have some similarities to classical topics such as fitting mixture models (McCulloch and Searle, 2001; Jiang, 2007). The primary difference is that in such work, first, *every* data point should have been drawn from some linear model in the mixture; if some large subset of the data cannot be fit well by linear models, there is no guarantee that the model will identify a small subset that can be fit well. A second difference is that such models do not provide a (DNF) rule to decide whether or not subsequent data is drawn from one of the components versus another. There are a number of topics such as regression trees (Quinlan, 1992), cluster-wise regression (Park et al., 2017), etc. that do provide such rules, but again, if the data overall cannot be fit well, they do not guarantee that small subsets of the data that can be fit well will be found.

2 PRELIMINARIES

We will assume that we have a data set containing N samples, from a distribution $\mathcal{D}$ over $\{0,1\}^n \times \mathbb{R}^d \times \mathbb{R}$. Each sample consists of an n dimensional vector of Boolean attributes $\boldsymbol{x}$, a d dimensional real valued vector of predictor variables $\boldsymbol{y}$, and a real valued response z, which we would like to predict. We will denote the ith sample as $(\boldsymbol{x}, \boldsymbol{y}, z)^{(i)}$ and abbreviate it as $\boldsymbol{x}^{(i)}$ when there is no ambiguity.

For linear regression, we want to find a vector of coefficients $\boldsymbol{v}$ such that z can be predicted by $\langle \boldsymbol{v}, \boldsymbol{y} \rangle$. Typically, $\boldsymbol{v}$ is found using ordinary least squares, which minimizes the sum of $(\langle \boldsymbol{v}, \boldsymbol{y} \rangle - z)^2$ over all data points. However, since we are interested in cases where the majority of data cannot be fit, we want to find a subset of points described by condition $\boldsymbol{c}$ where there exists a good linear model. Similar to previous work, we will consider conditions represented by Disjunctive Normal Form (DNF) formulas; other natural families of formulas are either weaker or yield intractable problems (Juba, 2017). A *k-DNF* is defined to be a disjunction (OR) of *terms* where each term is a conjunction (AND) of no more than k attributes.

Throughout this paper, we will use $\|\cdot\|_F$ to denote the Frobenius norm and $\|\cdot\|_2$ to denote the ℓ_2-norm of a vector. We will also define $\mathcal{X}_I$ as the characteristic function where $\mathcal{X}_I(x) = 1$ if $x \in I$ and 0 otherwise. For brevity, we will use $[N] = \{n \in \mathbb{N} | 1 \leq n \leq N\}$. For a matrix M, $M \succeq 0$ denotes M is positive semidefinite. Finally, we will use Π to denote projection matrices.

Definition 2.1 (Conditional Linear Regression)**.** *Given a sample of N points, $(\boldsymbol{x}, \boldsymbol{y}, z)^{(i)}$, from a distribution $\mathcal{D}$ over $\{0,1\}^n \times \mathbb{R}^d \times \mathbb{R}$, the task of conditional linear regression is to find a k-DNF, $\boldsymbol{c}$, and linear predictor, $\boldsymbol{v} = (v_1 \dots v_d)^T$, such that, with high probability, $|\langle \boldsymbol{v}, \boldsymbol{y} \rangle - z|$ is bounded by ϵ when conditioned on $\boldsymbol{c}(\boldsymbol{x}) = 1$ and $\boldsymbol{c}(\boldsymbol{x}) = 1$ is satisfied by at least a μ fraction of the data.*

We will present an algorithm that finds $\boldsymbol{c}$ and $\boldsymbol{v}$ that is close to the optimal values of $\boldsymbol{c}^*$ and $\boldsymbol{v}^*$ given that the distribution of samples conditioned on $\boldsymbol{c}^*$ follows certain regularity conditions. Our algorithm is obtained by solving a *sum-of-squares relaxation* (Parrilo, 2000; Lasserre, 2001; Nesterov, 2000; Shor, 1987) of a polynomial optimization problem:

Definition 2.2 (Sum-of-squares relaxation)**.** *Given a system of polynomial inequalities for polynomials in $\mathbb{R}[x_1, \dots, x_n]$, $g_1(\boldsymbol{x}) \geq 0, \dots, g_r(\boldsymbol{x}) \geq 0$, $h_1(\boldsymbol{x}) = 0, \dots, h_s(\boldsymbol{x}) = 0$, the* degree-$\ell$ sum-of-squares relaxation *is the following semidefinite optimization problem. The set of program variables $\boldsymbol{u}$ is indexed by monomials over $x_1, \dots, x_n$ of total degree at most ℓ,*

with $u_{\boldsymbol{\alpha}}$ denoting the variable for the monomial with degree vector $\boldsymbol{\alpha} \in \mathbb{N}^n$. We define the degree-$\ell$ moment matrix $M_\ell(\boldsymbol{u})$ *indexed by $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^n$ with total degree at most $\ell/2$ to be $M_\ell(\boldsymbol{u})_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = u_{\boldsymbol{\alpha}+\boldsymbol{\beta}}$. For a "shift" polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ of degree t, letting $p_{\boldsymbol{\gamma}}$ denote the coefficient of the monomial $\boldsymbol{x}^{\boldsymbol{\gamma}}$ in p, the* degree-ℓ localizing matrix $M_\ell(p\boldsymbol{u})$ *is defined by $M_\ell(p\boldsymbol{u})_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = \sum_{\boldsymbol{\gamma}} p_{\boldsymbol{\gamma}} u_{\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}}$ for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of total degree at most $\ell/2 - t$. Now, the program has the constraints that $u_{\mathbf{0}} = 1$; $M_\ell(\boldsymbol{u}) \succeq 0$; $M_\ell(h_j \boldsymbol{u}) = 0$ for $j \in \{1, \ldots, s\}$; and $M_\ell(\prod_{j \in S} g_j \boldsymbol{u}) \succeq 0$ for $S \subseteq \{1, \ldots, r\}$ s.t. $\sum_{j \in S} \deg(g_j) \leq \ell$.*

Given a bound on the magnitudes of the values involved, the solutions to semidefinite programs can be approximated to arbitrary precision in polynomial time by various algorithms; the current state-of-the-art is due to Jiang et al. (2020).

A helpful interpretation of the sum-of-squares relaxation is that it defines a set of "pseudo-distributions" that relax the moments of probability distributions supported on solutions to the system of inequalities, with an associated "pseudo-expectation" operator defined on polynomials of degree up to ℓ (we borrow the presentation from Raghavendra and Yau (2020)):

Definition 2.3 (Pseudo-distribution (Barak et al., 2012))**.** *A level ℓ pseudo-distribution is a finitely-supported function $D : \mathbb{R}^n \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x) f(x)^2 \geq 0$ for every polynomial f of degree at most $\ell/2$.*

Definition 2.4 (Pseudo-expectation (Barak et al., 2012))**.** *The pseudo-expectation of a function f on $\mathbb{R}^d$ with respect to a pseudo-distribution D, denoted by $\tilde{\mathbb{E}}_{D(x)} = \sum_x D(x) f(x)$.*

A low-degree pseudo-distribution is generally *not* a probability distribution, and we generally cannot sample from it. The quality of a sum-of-squares relaxation is characterized by "sum-of-squares *proofs*"—under mild conditions, the bounds obtained by the relaxation of a given degree match the optimal bound that can be proved via a sum-of-squares proof (Parrilo, 2000; Lasserre, 2001; Nesterov, 2000; Shor, 1987):

Definition 2.5 (Sum of Squares proofs (Grigoriev and Vorobjov, 2002))**.** *Fix a set of polynomial inequalities $\mathcal{A} = \{g_i(x) \geq 0\}_{i \in [m]} \cup \{h_i(x) = 0\}_{i \in [m']}$ in variables $x_1, \ldots, x_n$. A sum-of-squares proof of $q(x) \geq 0$ is an identity of the form*

$$\left(1 + \sum_{k \in [m'']} d_k^2(x)\right) \cdot q(x) = \sum_{j \in [m''']} s_j^2(x) + \sum_{S \subseteq [m]} a_S^2(x) \cdot \prod_{i \in S} g_i(x) + \sum_{i \in [m']} b_i(x) h_i(x),$$

where $\{s_j(x)\}_{j \in [m''']}$, $\{a_S(x)\}_{S \subseteq [m]}$, $\{b_i(x)\}_{i \in [m']}$, and $\{d_k(x)\}_{k \in [m'']}$ are real polynomials. If the expressions in the identity have total degree at most D, we denote that $q(x) \geq 0$ has a degree-D sum-of-squares proof from $\mathcal{A}$ (where $x_1, \ldots, x_n$ are the indeterminates) by

$$\mathcal{A} \overset{x_1, \ldots, x_n}{\underset{D}{\vdash}} q(x) \geq 0.$$

Moreover, a recent technique for extracting solutions from the sum-of-squares relaxation has been to use *identifiability* (Raghavendra et al., 2019): roughly, if there is a sum-of-squares proof that a portion of the solution is determined up to small ℓ_2 error, then we can read that portion of the solution off directly from the degree-1 variables. Conceptually, our analysis will follow this approach. Towards such an analysis, we require that the data/noise possess some niceness properties in a form that is amenable to sum-of-squares:

Definition 2.6 (Certifiable Hypercontractivity, Section 2.1 of Bakshi and Kothari (2021))**.** *A distribution $\mathcal{D}$ over $\mathbb{R}^d$ has C-hypercontractive degree-2 polynomials if for every $d \times d$ matrix Q,*

$$\mathbb{E}_{x \sim \mathcal{D}} \left(\boldsymbol{x}^\top Q \boldsymbol{x} - tr(Q)\right)^{2h} \leq (Ch)^h \left(\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left(\boldsymbol{x}^\top Q \boldsymbol{x}\right)^2\right)^h.$$

*We say that the hypercontractivity is ℓ-*certifiable *if there is a degree-ℓ sum-of-squares proof of this inequality with Q as an indeterminate.*

Lemma 2.7 (Certifiable Hypercontractivity Under Sampling. Lemma 6.11 in Bakshi and Kothari (2021))**.** *Let $\mathcal{D}$ be a 1-subgaussian, $2h$-certifiably C-hypercontractive distribution over $\mathbb{R}^d$. Let $\mathcal{S}$ be a set of $n = \Omega((hd)^{8h})$ i.i.d. samples from $\mathcal{D}$. Then, with probability at least $1 - 1/poly(n)$, the uniform distribution on $\mathcal{S}$ is h-certifiably $(2C)$-hypercontractive.*

In linear regression, it is standard and convenient to assume Gaussian residuals with mean 0 and constant variance σ^2. We stress that in order to empirically estimate the squared error of candidate linear models, we need some control on the moments of the loss function, which for various candidates corresponds to various quadratic polynomials; hypercontractivity of such quadratic polynomials gives such bounds. Many distributions, including the Gaussian distribution, uniform distribution on the Boolean hypercube, and log-concave distributions are examples of certifiably hypercontractive distributions; affine transformations, (bounded-weight) mixtures, and products of such distributions are also certifiably hypercontractive (Klivans et al., 2018; Bakshi and Kothari, 2021). It is thus reasonable to assume that the distribution is, in particular, certifiably hypercontractive. Lemma 2.7 shows that empirical distributions from a certifiably hypercontractive distribution are also certifiably

hypercontractive. We also assume that degree-2 polynomials similarly have *certifiably* bounded variances, which also holds for such distributions (Bakshi and Kothari, 2021).

3 RESULTS

3.1 Preprocessing

For our algorithm, we will consider the data as m disjoint subsets which we will call terms, $\{I_j\}_{j=1}^m$. To find k-DNF conditions, we will create a term for each setting of each set of k distinct attributes, but the algorithm can be used with any family of sets we choose. We will weight each term by the number of points, $|I_j|$, in term I_j. We will define I_{good} as the collection terms of the optimal k-DNF $\boldsymbol{c}^*$. From the perspective of the subspace recovery task (Bakshi and Kothari, 2021), I_{good} represents the collection of inliers, and the remaining points represent the outliers. Since the condition $\boldsymbol{c}^*$ is satisfied by a μ fraction of the data, $|I_{good}| = \mu N$. Additionally, our algorithm will assume that each term I_j is pairwise disjoint. This can be achieved by duplicating points that satisfy more than one term (Calderon et al., 2020). We have N data points and m terms, so there will be at most mN data points following the duplication procedure. Previously, $|I_{good}| = |\bigcup_{I_j \in I_{good}} I_j|$ which increases to $|I_{good}| = \sum_{I_j \in I_{good}} |I_j|$ with duplicate points. Therefore size of I_{good} may blow up at most by a factor of the number of terms in I_{good}. If we use N' to denote the number of points after duplication, notice that $N'_{good}/N' \geq N_{good}/mN$. Thus the proportion of good points after duplication is at least a μ/m (Calderon et al., 2020). Hereafter, N, I_{good}, $\{I_j\}_{j=1}^m$, and μ will be used to describe the data after duplication.

Our algorithm only obtains good estimates when each term in I_{good} is large. Since the contribution of each term is weighted by its size, we can remove sufficiently small terms without compromising the quality of the estimates. Finally, we will extend $\boldsymbol{y}^{(i)}$ by a constant, 1, to allow for an intercept in the linear model.

3.2 Main algorithm: identifying the data subspace

Recall that our goal is to find the linear predictor $\boldsymbol{v}$ such that $\left\langle \boldsymbol{v}, \boldsymbol{y}^{(i)} \right\rangle = z^{(i)}$ for all points satisfying condition $\boldsymbol{c}^*$. If we extend $\boldsymbol{v}$ with -1 and $\boldsymbol{y}^{(i)}$ with $z^{(i)}$, the previous equation is equivalent to $\left\langle \boldsymbol{v}, \boldsymbol{y}^{(i)} \right\rangle = 0$. This equation describes a hyperplane, which allows us to look at our conditional linear regression problem from the perspective of a subspace recovery problem. Note that both $\boldsymbol{v}$ and $\boldsymbol{y}^{(i)}$ are now $(d+2)$-dimensional vectors. Let us use Π to denote a projection matrix that projects onto this subspace and Π_* to denote the optimal projection.

Definition 3.1 (Reformulation of the Problem)**.** *Given a distribution, $\mathcal{D}$, over points $\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, z^{(i)}\}_{i=1}^N$ and predefined disjoint subsets, $\{I_j\}_{j=1}^m$. Let I_{good} be an unknown collection of terms containing points that satisfy $\boldsymbol{c}^*$ such that $P(\boldsymbol{x} \in I_{good}) \geq \mu$. If there exists a linear predictor, $\boldsymbol{v}^*$, such that $|\langle \boldsymbol{v}^*, \boldsymbol{y} \rangle| \leq \epsilon$ and a projection matrix, Π_*, that projects onto the hyperplane described by the linear predictor, then we want to find $\hat{\Pi}$ that approximates Π_* such that $\|\hat{\Pi} - \Pi_*\|_F \leq error$.*

Note that the Frobenius norm bounds the spectral norm from above, where the squared error of $\boldsymbol{v}$ is precisely $\boldsymbol{v}^\top \Pi_* \boldsymbol{v}$, so a solution to this reformulation gives an empirical estimate that is off by at most $\|\boldsymbol{v}\|_2^2 \cdot error$.

The advantage of this reformulation is that the subspace is uniquely determined by the set of terms, whereas if the subspace has lower dimension, there may be a large space of linear models, which are thus not identifiable: i.e., Π_* does not necessarily have rank $n-1$ because the data may lie on a lower-dimensional subspace, e.g., if not all covariates are informative. This matters because sum-of-squares relaxes a program capturing the *moments of a distribution over solutions*—if there is a unique solution Π_* to our estimation problem (if it is "identifiable"), then the first moments of Π in such a distribution must contain that unique solution, hence we may be able to read the values of Π_* off from a solution to the sum-of-squares relaxation. But, if the solution is not unique, the values obtained from the sum-of-squares program may be uninformative. What remains to be shown is that there is a fixed degree sum-of-squares *proof* that the solution is unique. Then, the values of a solution can indeed be read off from a solution of a corresponding degree sum-of-squares relaxation. (See Raghavendra et al. (2019) for an overview of this approach.)

One remaining issue is that even if the subspace containing the support of $\mathcal{D}|\mathbf{c}^*$ is identifiable once $\mathbf{c}^*$ is fixed, there may be multiple candidates $\mathbf{c}$ for $\mathbf{c}^*$. If each lies in a different subspace, the variables Π in a solution to the sum-of-squares relaxation may not correspond to any of these subspaces. An analogous issue arises in robust linear regression in the minority-inlier regime, and Karmalkar et al. (2019); Raghavendra and Yau (2020) developed techniques to address this issue: in particular, Algorithm 1 follows the "rounding by votes" approach of Karmalkar et al. (2019). In this approach, we (step 1) solve for, in particular, a set of moments that minimizes the ℓ_2-norm of the term indicator variables' first moments. Intuitively, this ensures that we obtain a maximum entropy mixture of such solutions, so in particular the solution has support on any desired $\mathbf{c}^*$ (c.f. Lemma 3.7). Then we (step 2) con-

sider the solutions we would obtain by *conditioning on* each one of these indicators. Intuitively, this "breaks ties," by asserting that one of the terms must be included in the program's solution. In particular, some term must supply at least an average fraction of the probability mass of $\mathbf{c}^*$, and by Markov's Inequality, the error from Π_* of this solution is only moderately larger than the error if we had conditioned on $\mathbf{c}^*$. This latter error is small (Lemma 3.6), and hence we obtain a sufficiently good estimate of Π_* by conditioning on a single term. In fact, we need not consider all of the indicators, (step 3) it suffices to consider the solutions obtained from a sample of $O(1/\mu)$ indicators, sampled according to their weight in the solution. Apart from the details of the relaxation, this is Algorithm 1.

The main issue is thus indeed to establish that the relaxation is sufficiently tight to permit identifiability of Π_* when conditioned on $\mathbf{c}^*$, as captured in Lemma 3.6. We remark that Karmalkar et al. (2019); Raghavendra and Yau (2020) needed a stronger assumption to obtain identifiability for regression, that the data was *anti-concentrated*. Anticoncentration does not hold if, e.g., the data lies on a lower-dimensional manifold, so it is a restrictive assumption. By contrast, we only need hypercontractivity and bounded moments to estimate the subspace.

We now state our main theorem, which is adapted from Theorem 1.4 of Bakshi and Kothari (2021).

Theorem 3.2. *Let Π_* be a projection matrix for a dimension r subspace. Let $\mathcal{D}|\boldsymbol{c}^*$ be a mean 0, covariance Π_* distribution with 2-certifiably C-hypercontracitve degree-2 polynomials with certifiably C-bounded variances. Then, there exists an algorithm that takes $n \geq \Omega\left((d \log d/\mu)^{16}\right)$ samples from $\mathcal{D}$ and outputs a list $\mathcal{L}$ of $\mathcal{O}(1/\mu)$ projection matrices such that with probability at least 0.99 over the draw of the sample and randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F \leq \mathcal{O}(1/\mu)$ in polynomial time.*

Let $\mathcal{A}_{w,v,\epsilon,\Pi}$ be the SoS program in Figure 1 where $\boldsymbol{w}$, $\boldsymbol{v}$, $\boldsymbol{\epsilon}$, and $\Pi_1, \ldots, \Pi_m$ are indeterminates.

We interpret the program constraints as follows:

1. Defines $\boldsymbol{y}'^{(i)}$ such that the inliers have mean 0 while preserving the constant 1 for the intercept of the model.
2. Conditioned on $\boldsymbol{c}$, the linear predictor $\boldsymbol{v}$ fits well.
3. The number of samples when conditioned on $\boldsymbol{c}$ comprises a μ fraction of the data.
4. The Boolean constraint: $w_j \in \{0, 1\}$ for all j.
5. Defines Π_j as a projection matrix corresponding to the distribution of points in term I_j.
6. The residuals of the linear model follow a Gaussian distribution with mean 0 and standard deviation σ, bounding the average noise on the inliers.
7. The samples are certifiably hypercontractive.
8. Similarly, the variance is certifiably bounded.
9. The second moment of each predictor variable of samples satisfying $\boldsymbol{c}$ is bounded by α.
10. Finally, the fourth moment of each predictor variable of samples satisfying $\boldsymbol{c}$ is bounded by β.

We note that constraints 7 and 8 are infinite families of constraints. But since there are sum-of-squares proofs of these constraints, similar to Bakshi and Kothari (2021), we can use the quantifier elimination technique (Fleming et al., 2019, Section 4.3.4) to rewrite these as standard constraints. We note that we substitute the cubic polynomial $\boldsymbol{y}'^{\top} Q \boldsymbol{y}'$ for Q in the proof of the hypercontractive inequality, thus obtaining a degree-6 sum-of-squares proof from the original degree-2 proof.

Algorithm 1

Input: Sample $\mathcal{Y} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$ from $\mathcal{D}$.
Output: A list $\mathcal{L}$ of $O(1/\mu)$ projection matrices such that there exists $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F < \mathcal{O}(1/\mu)$.
Operation:

1. Find a degree 12 pseudo-distribution $\tilde{\mu}$ satisfying $\mathcal{A}_{w,v,\epsilon,\Pi}$ that minimizes $\sqrt{\sum_{j=1}^m w_j \sum_{\boldsymbol{x}^{(i)} \in I_j} \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})}$.
2. For each $i \in [N]$ such that $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{j=1}^m w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)\right] > 0$, let $\hat{\Pi}_i = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{j=1}^m w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)\Pi\right]}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{j=1}^m w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)\right]}$. Otherwise, set $\hat{\Pi}_i = 0$.
3. Take J to be a random multi-set formed by union of $O(1/\mu)$ independent draws of $i \in [N]$ with probability $\frac{\tilde{\mathbb{E}}\left[\sum_{j=1}^m w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)\right]}{\mu N}$
4. Output $L = \{\hat{\Pi}_i | i \in J\}$ where $J \subseteq [N]$

For the following analysis, $\boldsymbol{y}^{(i)}$ will denote the samples after centering the inliers, which is equivalent to $\boldsymbol{y}'^{(i)}$ from the program.

The root of our improved analysis is Lemma 3.3 below, which uses Chebyshev's Inequality to give us an error bound inversely proportional to N^2. Thus, as we draw more data, the right hand side goes to 0.

Lemma 3.3 (Frobenius Closeness of Empirical and True Covariances)**.** *Define $w(I_j) = \frac{1}{\mu N}\sum_{i=1}^N w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})$. With probability $1 - \delta$,*

$$\left\| \frac{1}{\mu N} \sum_{i=1}^N w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \Pi_j \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)\top} \Pi_j - w(I_j) \Pi_j \right\|_F^2 \leq \frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4 \sigma^2)}{\mu^2 N^2 \delta}.$$

$\mathcal{A}_{w,v,\epsilon,\Pi}$:

$$\left\{\begin{array}{lr}
 & \boldsymbol{y}'^{(i)} = \boldsymbol{y}^{(i)} - \frac{1}{\mu N}\sum_{j=1}^{m}\sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\boldsymbol{y}^{(i)} + \begin{bmatrix}1\\ \mathbf{0}_{d+1}\end{bmatrix} \\
\forall j \in [m] & w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\left(\left\langle \boldsymbol{v}, \boldsymbol{y}'^{(i)}\right\rangle + \epsilon_i\right) = 0 \\
\forall j \in [m] & \sum_{j=1}^{m} |I_j| w_j = \mu N \\
\forall j \in [m] & w_j = w_j^2 \\
\forall j \in [m] & w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j\left(\boldsymbol{y}'^{(i)} - \begin{bmatrix}\mathbf{0}_{d+1}\\ \epsilon_i\end{bmatrix}\right) = w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\left(\boldsymbol{y}'^{(i)} - \begin{bmatrix}\mathbf{0}_{d+1}\\ \epsilon_i\end{bmatrix}\right) \\
\forall j \in [m] & \sum_{i\in I_j} w_j \epsilon_i \le w_j \frac{\sigma}{|I_j|} \\
\forall Q, j \in [m] & \frac{1}{\mu N}\sum_{i}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\left(\boldsymbol{y}'^{(i)\top} Q \boldsymbol{y}'^{(i)} - \mathrm{tr}(Q\Pi_j)\right)^2 \le \frac{C w_j}{\mu N}\sum_{i}^{N} \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}'^{(i)\top} Q \boldsymbol{y}'^{(i)})^2 \\
\forall Q, j \in [m] & \frac{1}{\mu N}\sum_{i}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}'^{(i)\top} Q \boldsymbol{y}'^{(i)})^2 \le C\|\Pi_j Q \Pi_j\|_F^2 \\
\forall j \in [m] & w_j \sum_{i\in I_j}\left[(\boldsymbol{y}_1'^{(i)})^2 \quad \dots \quad (\boldsymbol{y}_{d+2}'^{(i)})^2\right]^\top \le w_j |I_j| \alpha \mathbf{1} \\
\forall j \in [m] & w_j \sum_{i\in I_j}\left[(\boldsymbol{y}_1'^{(i)})^4 \quad \dots \quad (\boldsymbol{y}_{d+2}'^{(i)})^4\right]^\top \le w_j |I_j| \beta \mathbf{1}
\end{array}\right\}$$

Figure 1: Polynomial constraints used in the SoS Program

Proof. The squared Frobenius norm is equivalent to summing the square of each element. Thus we will bound the square of each element using Chebyshev's inequality and then compute the sum. The matrix $\frac{1}{\mu N}\sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j \boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top}\Pi_j$ can treated as a random quantity that represents an empirical estimate of $w(I_j)\Pi_j$.

Recall $w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j(\boldsymbol{y}^{(i)} - \boldsymbol{\epsilon}_i) = w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}^{(i)} - \boldsymbol{\epsilon}_i)$ from the program. Thus $w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j \boldsymbol{y}^{(i)} = w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)$. The empirical covariance matrix can be rewritten as $\frac{1}{\mu N}\sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)^\top$. Let us use the notation $A_{r,s}$ to denote the element in row r and column s of a matrix A. Using Chebyshev's Inequality, the square of each element in the Frobenius norm can be bounded by $\frac{w_j(d+2)^2}{\mu^2 N^2 \delta}\Big(\sum_{i=1}^{N} \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) Var(((\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)^\top)_{r,s})\Big)$ with probability $1 - \delta/(d+2)^2$.

Since $\boldsymbol{y}^{(i)}$ has been extended with one, let us treat the first element of $\boldsymbol{y}^{(i)}$ as 1. Therefore, when $r = 1$ or $s = 1$, $Var(((\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)^\top)_{r,s}) \le \alpha + \alpha^2\sigma^2$. For all other pairs of r and s, the variance is bounded by $\beta + 1\alpha^4$. Due to the same coordinate of $\boldsymbol{y}^{(i)}$ being 1 for all $i \in [N]$, the second moment, α, must be at least 1. Therefore, $Var(((\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)^\top)_{r,s}) \le \beta + \alpha + 7\alpha^4\sigma^2$.

Tying it all together, the square of each element can be bounded by $\frac{w_j|I_j|w_j(d+2)^2}{\mu^2 N^2 \delta}\left(\beta + \alpha + 7\alpha^4\sigma^2\right)$ with probability $1 - \delta/(d+2)^2$. This is a $(d+2)$ dimensional square matrix so there are $(d+2)^2$ elements. Therefore, the squared Frobenius norm is bounded by $\frac{w_j|I_j|w_j(d+2)^4}{\mu^2 N^2 \delta}\left(\beta + \alpha + 7\alpha^4\sigma^2\right)$ and by a union bound, this holds with probability at least $1 - \delta$. □

The improved bound of Lemma 3.3 then gives us an adequate estimate of the covariances for regression. We will crucially exploit the number of terms m being independent of the size of the sample N—this is the key difference between conditional linear regression and subspace recovery or robust linear regression.

Lemma 3.4 (Frobenius Closeness of Subsample to

Covariance, w-samples)**.**

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left\| \frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)\top} - w(I_j)\Pi_j \right\|_F^4 \right.$$
$$\leq C^2 \left(\frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \right)$$
$$\left. \cdot \left(\frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4 \sigma^2)}{\mu^2 N^2 \delta} \right) \right\}$$

with probability at least $1 - \delta$*.*

Lemma 3.5 (Frobenius Closeness of Subsample to Covariance, I-samples)**.**

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left\| \frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)\top} - w(I_j)\Pi_{j*} \right\|_F^4 \right.$$
$$\leq C^2 \left(\frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \right)$$
$$\left. \cdot \left(\frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4 \sigma^2)}{\mu^2 N^2 \delta} \right) \right\}$$

with probability at least $1 - \delta$*.*

The proofs of Lemmas 3.4 and 3.5 are similar to Lemmas 4.5 and 4.6 in Bakshi and Kothari (2021), but by using Lemma 3.3 we are able to form a tighter bound.

Lemma 3.6 (Frobenius Closeness of Π and Π_*, Lemma 4.3 in Bakshi and Kothari (2021))**.**

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left(\sum_{j=1}^{m} w(I_j) \right) \|\Pi - \Pi_*\|_F^2 \right.$$
$$\left. \leq mC \sqrt{2^5 \left(\frac{w_j (d+2)^4 (\beta + \alpha + 7\alpha^4 \sigma^2)}{\mu N \delta} \right)} \right\}.$$

with probability at least $1 - m\delta$ *where* $w(I_j) = \frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)$*.*

Lemma 3.6 admits a sum-of-squares proof and is proved by using Lemmas 3.4 and 3.5. Here, it is crucial that we can take $N \gg m$. The full proofs for all new lemmas are included in the supplementary material.

Lemma 3.7 (Large weight on inliers from high--entropy constraints. Fact 4.4 in Bakshi and Kothari (2021) and Lemma 3.1 in Raghavendra and Yau (2020))**.** *Let* $\tilde{\mathbb{E}}_\xi$ *be a pseudo-distribution of degree* ≥ 2 *that satisfies* $\mathcal{A}_{w,v,\epsilon,\Pi}$ *and minimizes* $\left\| \tilde{\mathbb{E}}_\xi \sum_{j=1}^{m} \sum_{i \in I_j} w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right) \right\|_2$*. Then* $\tilde{\mathbb{E}}_\xi \left[\sum_{j=1}^{m} \sum_{i \in I_j} w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right) \right] \geq \mu^2 N$*.*

The main theorem, Theorem 3.2, can be obtained by using Lemmas 3.6 and 3.7 and following a similar argument to Theorem 1.4 of Bakshi and Kothari (2021).

Proof. Given a distribution $\mathcal{D}|\mathbf{c}^*$ that is certifiably hypercontractive, Lemma 2.7 implies that a large enough sample of inliers will also be certifiably hypercontractive with high probability. Algorithm 1 finds a pseudo-distribution that satisfies $\mathcal{A}_{w,v,\epsilon,\Pi}$ and minimizes $\sqrt{\sum_{j=1}^{m} w_j \sum_{\boldsymbol{x}^{(i)} \in I_j} \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})}$. Then $\frac{1}{\mu N} \sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \|\Pi - \Pi_*\|_F^2 \right] \leq mC \sqrt{2^5 \frac{w_j (d+2)^4 (\beta+\alpha+7\alpha^4\sigma^2)}{\mu N \delta}}$ by using Lemma 3.6. By applying Jensen's Inequality and taking the square root, we have $\frac{1}{\mu N} \sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}} \left\| [\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi] - [\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_*] \right\|_F^2 \leq \sqrt{mC \sqrt{2^5 \frac{w_j (d+2)^4 (\beta+\alpha+7\alpha^4\sigma^2)}{\mu N \delta}}}$. Due to the definition of $\hat{\Pi}_i$ from the algorithm, we can rewrite the inequality as $\frac{1}{\mu N} \sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}} \|\hat{\Pi}_i - \Pi_*\|_F \leq \sqrt{mC \sqrt{2^5 \frac{w_j (d+2)^4 (\beta+\alpha+7\alpha^4\sigma^2)}{\mu N \delta}}}$. Let $Z = \frac{1}{\mu N} \sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})]$. Then, from Lemma 3.7, $Z \geq \mu$ and $\frac{1}{Z} \leq \frac{1}{\mu}$. Dividing by Z on both sides thus yields: $\frac{1}{Z} \left(\frac{1}{\mu N} \sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}} \|\hat{\Pi}_i - \Pi_*\|_F \right) \leq \frac{1}{\mu} \sqrt{mC \sqrt{2^5 \frac{w_j (d+2)^4 (\beta+\alpha+7\alpha^4\sigma^2)}{\mu N \delta}}}$.

Since each index $i \in [N]$ is chosen with probability $\widetilde{\mathbb{E}}[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}_i)] / \sum_{i \in [N]} \widetilde{\mathbb{E}}[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}_i)] = \frac{1}{\mu n} \widetilde{\mathbb{E}}[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}_i)]$, it follows that $i \in I_{good}$ with probability at least $\frac{1}{\mu n} \sum_{j=1}^{m} w_j \sum_{i \in I_j} \widetilde{\mathbb{E}}[\mathcal{X}_{I_j}(\boldsymbol{x}_i)] = Z \geq \mu$. By Markov's inequality applied to the last equation, with probability $\frac{1}{2}$ over the choice of i conditioned on $i \in I_{good}$, $\|\hat{\Pi}_i - \Pi_*\|_F \leq \frac{2}{\mu} \sqrt{mC \sqrt{2^5 \frac{w_j (d+2)^4 (\beta+\alpha+7\alpha^4\sigma^2)}{\mu N \delta}}}$. Thus, in total, with probability at least $\mu/2$, $\|\hat{\Pi}_i - \Pi_*\|_F \leq \frac{2}{\mu} \sqrt{mC \sqrt{2^5 \frac{w_j (d+2)^4 (\beta+\alpha+7\alpha^4\sigma^2)}{\mu N \delta}}}$. Therefore, with probability of at least 0.99 over the draw of the random set J, the list constructed by the algorithm contains $\hat{\Pi}$ such that $\|\hat{\Pi}_i - \Pi_*\|_F \leq \frac{2}{\mu} \sqrt{mC \sqrt{2^5 \frac{w_j (d+2)^4 (\beta+\alpha+7\alpha^4\sigma^2)}{\mu N \delta}}}$.

For the running time, the SDP can be solved in polynomial time and dominates the running time. Therefore, the algorithm runs in polynomial time overall. □

For a projection Π and a linear predictor, $\boldsymbol{v}$, which satisfies $\langle \boldsymbol{v}, \boldsymbol{y}^{(i)} \rangle = 0$ for all $i \in I_{good}$ when disregarding noise, $\boldsymbol{v}^\top \Pi = \mathbf{0}^\top$. Therefore, for each candidate Π, we can recover a candidate linear predictor $\boldsymbol{u}$ by treating $\boldsymbol{u}$ as a solution to the linear system $\boldsymbol{u}^\top \Pi = \mathbf{0}^\top$.

3.3 Obtaining a k-DNF Condition

Once we obtain an approximation for $\hat{\boldsymbol{v}}$, we will use the method described by Calderon et al. (2020) to obtain a k-DNF condition for $\boldsymbol{c}$. Since we have used the same reductions, specifically the reduction to disjoint terms by duplicating points, we will be able to invoke their analysis of their method. Let us define a loss function, $f^{(i)} : \mathcal{H} \to \mathbb{R}$, for each point as $f^{(i)}(\boldsymbol{v}) = \left(z^{(i)} - \left\langle \boldsymbol{v}, \boldsymbol{y}^{(i)} \right\rangle\right)^2$. Let $f_{I_j}(\boldsymbol{v})$ be a the average loss over points in I_j. Finally, let us define $\bar{f}(\boldsymbol{v}) = \mathbb{E}[f_{I_{good}}(\boldsymbol{v})]$ as the expectation of average loss for points in I_{good}. We will search through the Boolean data space $\{\boldsymbol{x}\}$ for conditions $\boldsymbol{c}$ for each $\boldsymbol{u}$. When we find a pair $(\boldsymbol{u}, \boldsymbol{c})$ such that $f_c(\boldsymbol{u})$ is small and there are are enough points satisfying $\boldsymbol{c}$, then we return that pair as a solution. For some constant accuracy parameter γ and Lipschitz constant L, if $\|\boldsymbol{u} - \boldsymbol{v}^*\| < \gamma$, then $|\bar{f}(\boldsymbol{u}) - \bar{f}(\boldsymbol{v}^*)\| \leq \gamma L = \mathcal{O}(\gamma)$. Since $\bar{f}$ is nonnegative, if $\bar{f}(\boldsymbol{v}^*) \leq \epsilon$, then $\bar{f}(\boldsymbol{u}) \leq \epsilon + \gamma$.

Recall that during preprocessing, we duplicated points that satisfied more than 1 term so that each term is disjoint. For m terms, each point can be copied at most m times. Let t be the number of terms in $\boldsymbol{c}$.

Lemma 3.8 (Lemma 3.4 in Calderon et al. (2020))**.** *Let $\boldsymbol{u}$ be such that $\|\boldsymbol{u} - \boldsymbol{v}^*\| < \gamma$. Then $|\bar{f}(\boldsymbol{u})| \leq t(\gamma + \epsilon)$.*

After obtaining $\boldsymbol{u}$ such that $f_i(\boldsymbol{u})$ is close to $f_i(\boldsymbol{v}^*)$, Calderon et al. (2020) use a greedy set-cover algorithm to find the corresponding conditions $\boldsymbol{c}$. The algorithm greedily chooses terms I_j satisfying $\sum_{i \in I_j} f^{(i)}(\boldsymbol{u}) \leq (1 + \gamma)\mu\epsilon N$ to maximize the number of additional points in I_j that did not satisfy the previously chosen terms. It iterates until the number of points satisfying the chosen terms is at least $(1 - \gamma/2)\mu N$.

Lemma 3.9 (Lemma 3.5 in Calderon et al. (2020))**.** *If there exists an optimal k-DNF $\boldsymbol{c}^*$ that is satisfied by a μ-fraction of the points with total loss ϵ, then, the weighted greedy set cover algorithm finds a k-DNF $\hat{\boldsymbol{c}}$ that is satisfied by a $(1-\gamma)\mu$-fraction of the points with total loss $\mathcal{O}(t \log(\mu N)\epsilon)$.*

Thus, we can obtain a pair $(\boldsymbol{u}, \boldsymbol{c})$ that gives empirical error that is only greater than the optimal by a $\mathcal{O}(t \log(\mu N))$ factor. Given that our assumed bounds on the moments of the data distribution implies that the square of the loss (being a quadratic polynomial) is bounded, we can use the bounds of Cortes et al. (2013) to bound the generalization error of linear regression on each possible k-DNF, to thus obtain that the true generalization error is similarly bounded. For $\boldsymbol{y} \in \mathcal{B} \subseteq \mathbb{R}^d$, where B is the ℓ_2 radius of $\mathcal{B}$, this only incurs a polynomial increase in the sample complexity in B, t, and $1/\gamma$ overall.

4 DISCUSSION AND FUTURE DIRECTIONS

We have thus shown that the assumption of homogeneous covariances used by Calderon et al. (2020) is not needed to obtain a polynomial-time algorithm for conditional linear regression. On the other hand, although we obtain a polynomial running time and sample complexity, the exponents are quite large. In particular, the sample complexity we obtain in our analysis is far from optimal for this problem. Thus, our algorithm is impractical in its current form and our contribution is strictly theoretical. The main direction for future work is to develop a practical algorithm that does not require the homogeneous covariance assumption.

We now elaborate on the obstacles. Much of the overhead in the sample complexity arises from the use of the certifiable hypercontractivity assumption (specifically Lemma 2.7, which guarantees certifiability is preserved for empirical distributions), which requires relatively high-degree polynomial expressions. Another source of sub-optimality seems to arise from the way we invoke bounds obtained via Chebyshev's inequality on the Frobenius error of the projector onto the data subspace. We conjecture that this can be improved by a more refined analysis, but it is still not clear whether or not certifiable hypercontractivity – the dominant source of overhead for most purposes – is really necessary.

The overhead in the computational complexity arises from the use of the sum-of-squares semidefinite program relaxation. Although the degree of the relaxation is moderate, the technique unfortunately yields algorithms that solve a large semidefinite program, and thus inherently tends to simply scale poorly. In many cases, however, the development of a sum-of-squares algorithm has led to the subsequent development of a spectral algorithm that can be practical. Examples of this sequence include a number of robust estimation tasks (Schramm and Steurer, 2017; Hopkins et al., 2019; Diakonikolas et al., 2020; Depersin, 2020) and tasks to identify hidden structures in a data set (Hopkins et al., 2016, 2017), that are variously of similar flavor to our problem.

Acknowledgements

This research is partially supported by NSF awards IIS-1908287, IIS-1939677, and CCF-1718380.

References

Bakshi, A. and P. K. Kothari (2021): "List-Decodable Subspace Recovery: Dimension Independent Error in Polynomial Time," in *Proceedings of*

the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, 1279–1297.

Barak, B., F. G. Brandao, A. W. Harrow, J. Kelner, D. Steurer, and Y. Zhou (2012): "Hypercontractivity, sum-of-squares proofs, and their applications," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 307–326.

Calderon, D., B. Juba, S. Li, Z. Li, and L. Ruan (2020): "Conditional linear regression," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2164–2173.

Charikar, M., J. Steinhardt, and G. Valiant (2017): "Learning from Untrusted Data," in *Proceedings of the 49th ACM Symposium on Theory of Computing*, 47–60.

Cortes, C., G. DeSalvo, and M. Mohri (2016): "Learning with rejection," in *International Conference on Algorithmic Learning Theory*, Springer, 67–82.

Cortes, C., S. Greenberg, and M. Mohri (2013): "Relative deviation learning bounds and generalization with unbounded loss functions," *arXiv preprint arXiv:1310.5796*.

Depersin, J. (2020): "A spectral algorithm for robust regression with subgaussian rates," *arXiv preprint arXiv:2007.06072*.

Diakonikolas, I., G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart (2019): "Robust estimators in high-dimensions without the computational intractability," *SIAM Journal on Computing*, 48, 742–864.

Diakonikolas, I., D. Kane, and D. Kongsgaard (2020): "List-decodable mean estimation via iterative multi-filtering," *Advances in Neural Information Processing Systems*, 33.

Diakonikolas, I., D. Kane, A. Pensia, T. Pittas, and A. Stewart (2021): "Statistical query lower bounds for list-decodable linear regression," in *Advances in Neural Information Processing Systems 34*, vol. 34.

El-Yaniv, R. and Y. Wiener (2012): "Pointwise tracking the optimal regression function," in *Advances in Neural Information Processing Systems 25*, 2042–2050.

Fischler, M. A. and R. C. Bolles (1981): "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, 24, 381–395.

Fleming, N., P. Kothari, and T. Pitassi (2019): "Semialgebraic Proofs and Efficient Algorithm Design," *Foundations and Trends® in Theoretical Computer Science*, 14, 1–221.

Grigoriev, D. and N. Vorobjov (2002): "Complexity of Null- and Positivstellensatz proofs," *Ann. Pure and Applied Logic*, 113, 153–160.

Hainline, J., B. Juba, H. S. Le, and D. P. Woodruff (2019): "Conditional sparse ℓ_p regression with optimal probability," in *Proc. 22nd AISTATS*, vol. 89 of *PMLR*, 369–382.

Hopkins, S. B., P. K. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer (2017): "The power of sum-of-squares for detecting hidden structures," in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 720–731.

Hopkins, S. B., T. Schramm, and J. Shi (2019): "A robust spectral algorithm for overcomplete tensor decomposition," in *Conference on Learning Theory*, PMLR, 1683–1722.

Hopkins, S. B., T. Schramm, J. Shi, and D. Steurer (2016): "Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 178–191.

Huber, P. J. (1981): *Robust Statistics*, New York, NY: John Wiley & Sons.

Jiang, H., T. Kathuria, Y. T. Lee, S. Padmanabhan, and Z. Song (2020): "A faster interior point method for semidefinite programming," in *Proceedings of the 61st Annual IEEE Symposium on the Foundations of Computer Science*.

Jiang, J. (2007): *Linear and Generalized Linear Mixed Models and Their Applications*, Berlin: Springer.

Juba, B. (2017): "Conditional Sparse Linear Regression," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, vol. 67, 45.

Karmalkar, S., P. Kothari, and A. Klivans (2019): "List-Decodable Linear Regression," in *Advances in Neural Information Processing Systems 32*, 7423–7432.

Klivans, A., P. K. Kothari, and R. Meka (2018): "Efficient Algorithms for Outlier-Robust Regression," in *Proceedings of the 31st Conference On Learning Theory*, vol. 75 of *PMLR*, 1420–1430.

Lasserre, J. B. (2001): "Global Optimzation with Polynomials and the Problem of Moments," *SIAM J. Optimization*, 11, 796–817.

McCulloch, C. E. and S. R. Searle (2001): *Generalized, Linear, and Mixed Models*, New York, NY: John Wiley & Sons.

Nesterov, Y. (2000): "Squared functional systems and optimization problems," *High performance optimization*, 13, 405–440.

Park, Y. W., Y. Jiang, D. Klabjan, and L. Williams (2017): "Algorithms for Generalized Cluster-wise Linear Regression," *INFORMS Journal on Computing*, 29, 301–317.

Parrilo, P. A. (2000): "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. thesis, California Institute of Technology.

Prasad, A., A. S. Suggala, S. Balakrishnan, and P. Ravikumar (2020): "Robust estimation via robust gradient estimation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 601–627.

Quinlan, J. R. (1992): "Learning with continuous classes," in *5th Australian Joint Conference on Artificial Intelligence*, Singapore, vol. 92, 343–348.

Raghavendra, P., T. Schramm, and D. Steurer (2019): "High dimension estimation via sum-of-squares proofs," in *Proceedings of the International Congress of Mathematicians (ICM 2018)*, 3389–3423.

Raghavendra, P. and M. Yau (2020): "List decodable learning via sum of squares," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 161–180.

Rosenfeld, A., D. G. Graham, R. Hamoudi, R. Butawan, V. Eneh, S. Kahn, H. Miah, M. Niranjan, and L. B. Lovat (2015): "MIAT: A Novel Attribute Selection Approach to Better Predict Upper Gastrointestinal Cancer," in *Proc. IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1–7.

Rousseeuw, P. J. and A. M. Leroy (1987): *Robust Regression and Outlier Detection*, New York, NY: John Wiley & Sons.

Schramm, T. and D. Steurer (2017): "Fast and robust tensor decomposition with applications to dictionary learning," in *Conference on Learning Theory*, PMLR, 1760–1793.

Shor, N. (1987): "An approach to obtaining global extremums in polynomial mathematical programming problems," *Cybernetics and Systems Analysis*, 23, 695–700.

Conditional Linear Regression for Heterogeneous Covariances: Supplementary Materials

A Sum of Squares

We first recall the notion of pseudodistributions satisfying a set of constraints:

Definition A.1. *Let D be a level-ℓ pseudodistribution and $\mathcal{A}$ be a system of polynomial inequality constraints of the form $f_i \geq 0$ for $i \in [m]$. We will say that D satisfies A at degree r if for every subset $S \subseteq [m]$ of the constraints of $\mathcal{A}$ and every sum-of-squares polynomial h with $\deg(h) + \sum_{i \in S} \max\{\deg(f_i), r\} \leq \ell$, $\tilde{\mathbb{E}}_D[h \cdot \prod_{i \in S} f_i] \geq 0$. We denote this by $D \models_r \mathcal{A}$.*

Lemma A.2 (Soundness. Fact 3.4 in Bakshi and Kothari (2021)/Lemma 3.4 in Ma et al. (2016))**.** *If $D \models_r \mathcal{A}$ for a level-ℓ. pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \vdash_{r'} \mathcal{B}$, then $D \models_{r \cdot r' + r'} \mathcal{B}$.*

In words, if a pseudodistribution D satisfies the set of constraints $\mathcal{A}$ and there is a proof of another set of constraints $\mathcal{B}$ from $\mathcal{A}$, then if the level ℓ of D is sufficiently high, D must also satisfy $\mathcal{B}$. Hence, simply by optimizing a program formulated using $\mathcal{A}$, we obtain a solution that obeys any set of derived constraints $\mathcal{B}$.

Lemma A.3 (SoS Hölder's Inequality. Fact 3.11 in Bakshi and Kothari (2021)/Fact A.6 in Hopkins and Li (2017))**.** *Let $w_1, \ldots, w_n$ be indeterminates and let $f_1, \ldots, f_n$ be polynomials of degree m in vector valued variable x. Let k be a power of 2. Then,*

$$\left\{w_i^2 = w_i, \forall i \in [n]\right\} \vdash_{2kn}^{x,w} \left\{ \left(\frac{1}{n}\sum_{i=1}^n w_i f_i\right)^k \leq \left(\frac{1}{n}\sum_{i=1}^n w_i\right)^{k-1} \left(\frac{1}{n}\sum_{i=1=}^n f_i^k\right) \right\}.$$

Lemma A.4 (SoS Almost Triangle Inequality. Fact 3.8 in Bakshi and Kothari (2021)/Lemma A.2 in Kothari and Steurer (2017))**.** *Let a, b be indetermnates. Then, for any $t \in \mathbb{N}$,*

$$\vdash_{2t}^{a,b} \left\{(a+b)^{2t} \leq 2^{2t}\left(a^{2t} + b^{2t}\right)\right\}.$$

Lemma A.5 (Cancellation within SoS. Fact 3.12 in Bakshi and Kothari (2021)/Fact 5.4 in De et al. (2016))**.** *Let a be an indeterminate. Then,*

$$\left\{a^t \leq 1\right\} \cup \left\{a \geq 0\right\} \vdash_t^a \left\{a \leq 1\right\}.$$

Lemma A.6 (SoS Cauchy Schwarz. Fact 2.4 in Raghavendra and Yau (2020)/Lemma A.1 in Ma et al. (2016))**.** *Let $x_1, \ldots, x_n, y_1, \ldots, y_n$ be indeterminates, then*

$$\vdash_4^{x_1,\ldots,x_n,y_1,\ldots,y_n} \left\{ \left(\sum_{i=1}^n x_i y_i\right)^2 \leq \left(\sum_{1=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right) \right\}.$$

Lemma A.7 (Simple SoS AM-GM Inequality.)**.** *Let $f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n)$ be degree-d polynomials. Then,*

$$\vdash_{2d}^{x_1,\ldots,x_n} \left\{ \left(\frac{f_1(x_1,\ldots,x_n) + f_2(x_1,\ldots,x_n)}{2}\right)^2 \geq f_1(x_1,\ldots,x_n) \cdot f_2(x_1,\ldots,x_n) \right\}.$$

Proof. $\left(\frac{f_1(x_1,\ldots,x_n)+f_2(x_1,\ldots,x_n)}{2}\right)^2 - f_1(x_1,\ldots,x_n) \cdot f_2(x_1,\ldots,x_n) = \left(\frac{f_1(x_1,\ldots,x_n)-f_2(x_1,\ldots,x_n)}{2}\right)^2$. □

B Analysis of Algorithm 1

Lemma B.1 (Lemma 3.7 – Large weight on inliers from high-entropy constraints. Fact 4.4 in Bakshi and Kothari (2021) and Lemma 3.1 in Raghavendra and Yau (2020))**.** *Let $\tilde{\mathbb{E}}_\xi$ be a pseudo-distribution of degree ≥ 2 that satisfies $\mathcal{A}_{w,v,\epsilon,\Pi}$ and minimizes $\left\|\tilde{\mathbb{E}}_\xi \sum_{j=1}^m \sum_{i\in I_j} w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)\right\|_2$. Then $\tilde{\mathbb{E}}_\xi \left[\sum_{j=1}^m \sum_{i\in I_j} w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)\right] \geq \mu^2 N$.*

Proof. (This proof is the same as Lemma 3.1 in Raghavendra and Yau (2020).) For the sake of of simplicity, let $w_i = \sum_{i\in I_j} w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)$ and note that $w_i \in \{0,1\}$. Let $\tilde{\mathbb{E}}_P$ denote a pseudo-distribution corresponding to the actual assignment $\{w_i'\}_{i\in[N]}$ and let $\tilde{\mathbb{E}}_D$ be the pseudo-expectation that minimizes $\|\mathbb{E}_D[w]\|$. For a constant $\kappa \in [0,1]$, define the pseudo-expectation $\tilde{\mathbb{E}}_R$ as a mixture of $\tilde{\mathbb{E}}_P$ and $\tilde{\mathbb{E}}_D$.

$$\tilde{\mathbb{E}}_R \stackrel{def}{=} \kappa\tilde{\mathbb{E}}_P + (1-\kappa)\tilde{\mathbb{E}}_D$$

Since $\mathbb{E}_D$ is the pseudo-expectation that minimizes $\|\tilde{\mathbb{E}}_D[w]\|$, then

$$\left\langle \tilde{\mathbb{E}}_R[w], \tilde{\mathbb{E}}_R[w] \right\rangle \geq \left\langle \tilde{\mathbb{E}}_D[w], \tilde{\mathbb{E}}_D[w] \right\rangle.$$

We can use the definition of $\tilde{\mathbb{E}}_R$ to expand the left hand side.

$$\kappa^2 \left\langle \tilde{\mathbb{E}}_P[w], \tilde{\mathbb{E}}_P[w] \right\rangle + 2\kappa(1-\kappa) \left\langle \tilde{\mathbb{E}}_P[w], \tilde{\mathbb{E}}_D[w] \right\rangle + (1-\kappa)^2 \left\langle \tilde{\mathbb{E}}_D[w], \tilde{\mathbb{E}}_D[w] \right\rangle \geq \left\langle \tilde{\mathbb{E}}_D[w], \tilde{\mathbb{E}}_D[w] \right\rangle$$

By rearranging the terms, we get

$$\left\langle \tilde{\mathbb{E}}_P[w], \tilde{\mathbb{E}}_D[w] \right\rangle \geq \frac{1}{2\kappa(1-\kappa)} \left((2\kappa - \kappa^2) \left\langle \tilde{\mathbb{E}}_D[w], \tilde{\mathbb{E}}_D[w] \right\rangle - \kappa^2 \left\langle \tilde{\mathbb{E}}_P[w], \tilde{\mathbb{E}}_P[w] \right\rangle \right).$$

By definition, $\left\langle \tilde{\mathbb{E}}_P[w], \tilde{\mathbb{E}}_P[w] \right\rangle = \sum_{i=1}^N w_i^2 = \mu N$. By using the Cauchy-Schwartz inequality, $\left\langle \tilde{\mathbb{E}}_D[w], \tilde{\mathbb{E}}_D[w] \right\rangle \geq \frac{1}{N}\left(\sum_i \tilde{\mathbb{E}}_D[w_i]\right)^2 = \frac{1}{N}(\mu N)^2 = \mu^2 N$. By substituting these bounds, we get that

$$\left\langle \tilde{\mathbb{E}}_D[w], \tilde{\mathbb{E}}_P[w] \right\rangle \geq \frac{(2\kappa - \kappa^2)\mu^2 - \kappa^2\mu}{2\kappa(1-\kappa)} \cdot N.$$

As $\kappa \to 0$, the right hand side tends to $\mu^2 N$. □

Lemma B.2 (Lemma 3.3 – Frobenius Closeness of Empirical and True Covariances)**.** *Let $\boldsymbol{\epsilon}_i = \begin{bmatrix} \mathbf{0}_{d+1} \\ \epsilon_i \end{bmatrix}$ and $w(I_j) = \frac{1}{\mu N}\sum_{i=1}^N w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})$. Then with probability $1-\delta$,*

$$\left\| \frac{1}{\mu N} \sum_{i=1}^N w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \Pi_j \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)\top} \Pi_j - w(I_j)\Pi_j \right\|_F^2 \leq \frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4\sigma^2)}{\mu^2 N^2 \delta}.$$

Proof. The squared Frobenius norm is equivalent to summing the square of each element. Thus we will bound the square of each element using Chebyshev's inequality and then compute the sum. The matrix $\frac{1}{\mu N}\sum_{i=1}^N w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \Pi_j \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)\top} \Pi_j$ can treated as a random quantity that represents an empirical estimate of $w(I_j)\Pi_j$.

Recall $w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j(\boldsymbol{y}^{(i)} - \boldsymbol{\epsilon}_i) = w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}^{(i)} - \boldsymbol{\epsilon}_i)$ from the program. Thus $w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j \boldsymbol{y}^{(i)} = w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)$. The empircal covariance matrix can be rewritten as $\frac{1}{\mu N}\sum_{i=1}^N w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)^\top$. Let us use the notation $A_{r,s}$ to denote the element in row r and column s of a matrix A. Using Chebyshev's Inequality, the square of each element in the Frobenius norm can be bounded by $\frac{w_j(d+2)^2}{\mu^2 N^2 \delta}\left(\sum_{i=1}^N \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) Var(((\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)^\top)_{r,s})\right)$ with probability $1 - \delta/(d+2)^2$.

$Var(((\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)} + (\Pi_j - I)\boldsymbol{\epsilon}_i)^\top)_{r,s})$ is the variance of a sum of random variables. $Var(\boldsymbol{y}_r^{(i)}\boldsymbol{y}_s^{(i)}) = E[\boldsymbol{y}_r^{(i)2}\boldsymbol{y}_s^{(i)2}] - E[\boldsymbol{y}_r^{(i)}\boldsymbol{y}_s^{(i)}]^2 \leq E[\boldsymbol{y}_r^{(i)2}\boldsymbol{y}_s^{(i)2}]$ due to nonnegativity. By using the SoS Cauchy Schwarz and SoS

AM-GM inequalities, Lemmas A.6 and A.7, $E[\boldsymbol{y}_r^{(i)2}\boldsymbol{y}_s^{(i)2}] \le \sqrt{E[\boldsymbol{y}_r^{(i)2}]E[\boldsymbol{y}_s^{(i)2}]} \le \left(E[\boldsymbol{y}_r^{(i)4}] + E[\boldsymbol{y}_s^{(i)4}]\right)/2$. Thus $Var(\boldsymbol{y}_r^{(i)}\boldsymbol{y}_s^{(i)}) \le \max\{E[\boldsymbol{y}_r^{(i)4}]E[\boldsymbol{y}_s^{(i)4}]\} \le \beta$. Since $\boldsymbol{y}^{(i)}$ has been extended with one, let us treat the first element of $\boldsymbol{y}^{(i)}$ as 1. Therefore, when $r = 1$ or $s = 1$, $Var(((\boldsymbol{y}^{(i)}+(\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)}+(\Pi_j - I)\boldsymbol{\epsilon}_i)^\top)_{r,s}) \le \alpha+\alpha^2\sigma^2$. For all other pairs of r and s, the variance is bounded by $\beta+1\alpha^4$. Due to the same coordinate of $\boldsymbol{y}^{(i)}$ being 1 for all $i \in [N]$, the second moment, α, must be at least 1. Therefore, $Var(((\boldsymbol{y}^{(i)}+(\Pi_j - I)\boldsymbol{\epsilon}_i)(\boldsymbol{y}^{(i)}+(\Pi_j - I)\boldsymbol{\epsilon}_i)^\top)_{r,s}) \le \beta+\alpha+7\alpha^4\sigma^2$.

Tying it all together, the square of each element can be bounded by $\frac{w_j|I_j|w_j(d+2)^2}{\mu^2N^2\delta}\left(\beta+\alpha+7\alpha^4\sigma^2\right)$ with probability $1-\delta/(d+2)^2$. This is a $(d+2)$ dimensional square matrix so there are $(d+2)^2$ elements. Therefore,

$$\left\|\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j\boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top}\Pi_j - w(I_j)\Pi_j\right\|_F^2 \le \frac{w_j|I_j|(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu^2N^2\delta}.$$

By taking the union bound, this holds with probability at least $1-\delta$. □

Lemma B.3 (Lemma 3.4 – Frobenius Closeness of Subsample to Covariance, w-samples. Same proof as Lemma 4.5 in Bakshi and Kothari (2021))**.**

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,\Pi}_{12} \left\{ \left\|\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top} - w(I_j)\Pi_j\right\|_F^4 \le C^2\left(\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\right)\left(\frac{w_j|I_j|(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu^2N^2\delta}\right)\right.$$

with probability at least $1-\delta$.

Proof. For a $(d+2)\times(d+2)$ matrix-valued indeterminate Q and using the SoS Hölder's Inequality, Lemma A.3, we have

$$\begin{aligned}
\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi,Q}_{12} \left\{ \left\langle \frac{1}{\mu N}\sum_{i\in I_j} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top} - w(I_j)\Pi_j, Q\right\rangle^2 \right.& \\
= \left\langle \frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\left(\boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top} - \Pi_j\right), Q\right\rangle^2 & \qquad (1)\\
\left. \le \left(\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\right)\left(\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\left\langle \boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top} - \Pi_j, Q\right\rangle^2\right) \right\} &
\end{aligned}$$

Using certifiable hypercontractivity combined with the bounded variance constraints, we have

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi,Q}_{12} \left\{ \frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\left\langle \boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top} - \Pi_j, Q\right\rangle^2 \le \left(C^2t\right)\|\Pi_jQ\Pi_j\|_F^2 \right\}. \qquad (2)$$

By combining Equations 1 and 2 and substituting $Q = \frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top} - w(I_j)\Pi$, we have

$$\begin{aligned}
&\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left\|\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top} - w(I_j)\Pi_j\right\|_F^4 \right.\\
&\le C^2\left(\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\right)\|\Pi_jQ\Pi_j\|_F^2\\
&\left. = C^2\left(\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\right)\left\|\frac{1}{\mu N}\sum_{i=1}^{N} w_j\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_j\boldsymbol{y}^{(i)}\boldsymbol{y}^{(i)\top}\Pi_j - w(I_j)\Pi_j\right\|_F^2 \right\}.
\end{aligned}$$

By using Lemma B.2, we have

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left\| \frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)\top} - w(I_j)\Pi_j \right\|_F^4 \right.$$
$$\left. \leq C^2 \left(\frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \right) \left(\frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4\sigma^2)}{\mu^2 N^2 \delta} \right) \right\}$$

with probability at least $1 - \delta$. □

Lemma B.4 (Lemma 3.5 – Frobenius Closeness of Subsample to Covariance, I-samples. Same proof as Lemma 4.5 in Bakshi and Kothari (2021))**.**

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left\| \frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)\top} - w(I_j)\Pi_{j*} \right\|_F^4 \right.$$
$$\left. \leq C^2 \left(\frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) \right) \left(\frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4\sigma^2)}{\mu^2 N^2 \delta} \right) \right\}$$

with probability at least $1 - \delta$*.*

Proof. This follows the same proof as Lemma B.3. □

Lemma B.5 (Lemma 3.6 – Frobenius Closeness of Π and Π_*. Same as Lemma 4.3 in Bakshi and Kothari (2021).)**.**

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left(\sum_{j=1}^{m} w(I_j) \right) \|\Pi - \Pi_*\|_F^2 \right.$$
$$\left. \leq mC \sqrt{2^5 \left(\frac{w_j (d+2)^4 (\beta + \alpha + 7\alpha^4\sigma^2)}{\mu N \delta} \right)} \right\}.$$

with probability at least $1 - m\delta$ *where* $w(I_j) = \frac{1}{\mu N} \sum_{i=1}^{N} w_j \mathcal{X}_{I_j}\left(\boldsymbol{x}^{(i)}\right)$*.*

Proof. Define $w^*(I_j) = \frac{|I_j|}{\mu N}$. Using the SoS Almost Triangle Inequality, Lemma A.4, and Lemmas B.3 and B.4, we have

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ w(I_j)^4 \|\Pi_j - \Pi_{j*}\|_F^4 \right.$$
$$\left. \leq 2^5 C^2 w(I_j) \left(\frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4\sigma^2)}{\mu^2 N^2 \delta} \right) \right\}.$$

By dividing both sides of the inequality by $w(I_j)^2$, we have

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ w(I_j)^2 \|\Pi_j - \Pi_{j*}\|_F^4 \right.$$
$$\left. \leq 2^5 C^2 w(I_j)^{-1} \left(\frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4\sigma^2)}{\mu^2 N^2 \delta} \right) \right\}.$$

By using Cancellation within SoS, Lemma A.5 and multiplying both sides of the inequality by $w(I_j)^{1/2}$, we have

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ w(I_j) w(I_j)^{1/2} \|\Pi_j - \Pi_{j*}\|_F^2 \right.$$
$$\left. \leq C \sqrt{2^5 \left(\frac{w_j |I_j| (d+2)^4 (\beta + \alpha + 7\alpha^4\sigma^2)}{\mu^2 N^2 \delta} \right)} \right\}.$$

Since $w(I_j) = \frac{1}{\mu N}\sum_{i=1}^{N} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})$ and is bounded above by $w^*(I_j)$, then with probability $1-2\delta$:

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ w(I_j) \left\|\Pi_j - \Pi_{j*}\right\|_F^2 \leq C\sqrt{2^5\left(\frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}\right)}\right\}.$$

Define Π and Π_* as a weighted average of Π_j and Π_{j*} respectively, where the weights are proportional to $|I_j|$. Thus $\Pi = \sum_{j=1}^{m} w(I_j)\Pi_j$ and $\Pi_* = \sum_{j=1}^{m} w(I_j)\Pi_{j*}$. By summing both sides of the inequality over all $j \in [m]$, we have

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \sum_{j=1}^{m} w(I_j) \left\|\Pi_j - \Pi_{j*}\right\|_F^2 \leq \sum_{j=1}^{m} C\sqrt{2^5\left(\frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}\right)}\right\}.$$

By using the Cauchy-Schwarz Inequality and the triangle inequality to rewrite the left hand side, we have

$$\mathcal{A}_{w,v,\epsilon,\Pi} \vdash^{w,v,\epsilon,\Pi}_{12} \left\{ \left(\sum_{j=1}^{m} w(I_j)\right) \left\|\Pi - \Pi_*\right\|_F^2 \leq mC\sqrt{2^5\left(\frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}\right)}\right\}$$

with probability $1-2m\delta$. □

Lemma B.6 (Lemma 2.7 – Certifiable Hypercontractivity Under Sampling. Lemma 6.11 in Bakshi and Kothari (2021))**.** *Let $\mathcal{D}$ be a 1-subgaussian, $2h$-certifiably C-hypercontractive distribution over $\mathbb{R}^d$. Let $\mathcal{S}$ be a set of $n = \Omega((hd)^{8h})$ i.i.d. samples from $\mathcal{D}$. Then, with probability at least $1 - 1/poly(n)$, the uniform distribution on $\mathcal{S}$ is h-certifiably $(2C)$-hypercontractive.*

B.1 Proof of Main Theorem

Theorem B.7 (Main Theorem – Theorem 3.2)**.** *Let Π_* be a projection matrix for a subspace of dimension r. Let $\mathcal{D}|\mathbf{c}^*$ be a distribution with mean 0, covariance Π_*, and 2-certifiably C-hypercontracitve degree-2 polynomials. Then, there exists an algorithm that takes $n \geq \Omega\left((d\log d/\mu)^{16}\right)$ samples from the distribution $\mathcal{D}$ and outputs a list $\mathcal{L}$ of $O(1/\mu)$ projection matrices such that with probability at least 0.99 over the draw of the sample and randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F \leq O(1/\mu)$ in polynomial time.*

Proof. This follows the same proof as Theorem 1.4 in Bakshi and Kothari (2021). Since $\mathcal{D}|\mathbf{c}^*$ is certifiably C-hypercontractive, Fact B.6 implies that $\geq n = \Omega(d\log d/\mu)^{16}$ samples suffice for the uniform distribtution on the inliers, I_{good}, to have 2-certifiably C-hypercontractive degree 2 polynomials with probabiliy at least $1-1/d$. Let ξ_1 be the event that this succeds, and condition on it.

Let $\tilde{\mu}$ be a pseudo-distribution of degree-24 satisfying $\mathcal{A}_{w,v,\epsilon,\Pi}$ and minimizing $\sqrt{\sum_{j=1}^{m} w_j \sum_{i=1}^{N} \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})}$ as described in Algorithm 1. Observe that such a pseudo-distribution is guaranteed to exist: take the pesudo-distribution supported on a single point, (w,Π) such that $w_i = 1$ iff $i \in I_{good}$ and $\Pi = \Pi_*$. It is straight forward to check that Π_* is indeed a rank r projection matrix and rank $d+2$ projection matrix and $\sum_{j=1}^{m}\sum_{i=1}^{N} \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)}) = \mu N$. Conditioned on ξ_1, the hypercontractivity constraint is also satisfied by the inliers.

Since Lemma B.5 admits a sum-of-squares proof, it follows from Fact A.2 that the polynomial inequality is preserved under pseudo-expectations.

$$\frac{1}{\mu M}\sum_{j}^{k} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\|\Pi - \Pi_*\|_F^2\right] \leq mC\sqrt{2^5\frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}.$$

Alternatively, we can rewrite the above as follows:

$$\frac{1}{\mu N}\sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}} \left\| [\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi] - [\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_*] \right\|_F^2 \leq mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}.$$

Applying Jensen's Inequality yields

$$\left(\frac{1}{\mu N}\sum_{j=1}^{m} w_j \sum_{i=1}^{N} \left\| \tilde{\mathbb{E}}_{\tilde{\mu}}[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi] - \tilde{\mathbb{E}}_{\tilde{\mu}}[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_*] \right\|_F \right)^2 \leq mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}.$$

Taking the square root,

$$\frac{1}{\mu N}\sum_{j=1}^{m} w_j \sum_{i=1}^{N} \left\| \tilde{\mathbb{E}}_{\tilde{\mu}}[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi] - \tilde{\mathbb{E}}_{\tilde{\mu}}[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi_*] \right\|_F \leq \sqrt{mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}}.$$

Recall, the rounding in Algorithm 1 uses $\hat{\Pi}_i = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\Pi\right]}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})\right]}$ to denote the projector corresponding to the i-th sample. Then rewriting the above equation yields:

$$\frac{1}{\mu N}\sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}} \|\hat{\Pi}_i - \Pi_*\|_F \leq \sqrt{mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}}.$$

Let $Z = \frac{1}{\mu N}\sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}[\mathcal{X}_{I_j}(\boldsymbol{x}^{(i)})]$. Then, from Lemma B.1, $Z \geq \mu \Rightarrow \frac{1}{Z} \leq \frac{1}{\mu}$. Dividing by Z on both sides thus yields:

$$\frac{1}{Z}\left(\frac{1}{\mu N}\sum_{j=1}^{m} w_j \sum_{i=1}^{N} \tilde{\mathbb{E}}_{\tilde{\mu}} \|\hat{\Pi}_i - \Pi_*\|_F \right) \leq \frac{1}{\mu}\sqrt{mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}}.$$

Since each index $i \in [N]$ is chosen with probability $\frac{\widetilde{\mathbb{E}}[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}_i)]}{\sum_{i\in[N]} \widetilde{\mathbb{E}}[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}_i)]} = \frac{1}{\mu N}\widetilde{\mathbb{E}}[\sum_{j=1}^{m} w_j \mathcal{X}_{I_j}(\boldsymbol{x}_i)]$, it follows that $i \in I_{good}$ with probability at least $\frac{1}{\mu N}\sum_{j=1}^{m} w_j \sum_{i\in I_j} \widetilde{\mathbb{E}}[\mathcal{X}_{I_j}(\boldsymbol{x}_i)] = Z \geq \mu$. By Markov's inequality applied to the last equation, with probability $\frac{1}{2}$ over the choice of i conditioned on $i \in I_{good}$, $\|\hat{\Pi}_i - \Pi_*\|_F \leq \frac{2}{\mu}\sqrt{mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}}$. Thus, in total, with probability at least $\mu/2$, $\|\hat{\Pi}_i - \Pi_*\|_F \leq \frac{2}{\mu}\sqrt{mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}}$. Therefore, with probability of at least 0.99 over the draw of the random set J, the list constructed by the algorithm contains $\hat{\Pi}$ such that $\|\hat{\Pi}_i - \Pi_*\|_F \leq \frac{2}{\mu}\sqrt{mC\sqrt{2^5 \frac{w_j(d+2)^4(\beta+\alpha+7\alpha^4\sigma^2)}{\mu N\delta}}}$.

Now to account for the running time of the algorithm, the SDP for the program can be solved in polynomial time, so the algorithm runs in polynomial time overall. □

References

Bakshi, A. and P. K. Kothari (2021): "List-Decodable Subspace Recovery: Dimension Independent Error in Polynomial Time," in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, SIAM, 1279–1297.

De, A., E. Mossel, and J. Neeman (2016): "Majority is Stablest: Discrete and SoS," *Theory of Computing*, 12, 1–50.

Hopkins, S. B. and J. Li (2017): "Mixture Models, Robustness, and Sum of Squares Proofs," arXiv:1711.07454, abstract appeared in STOC 2018.

Kothari, P. K. and D. Steurer (2017): "Outlier-robust moment-estimation via sum-of-squares," *arXiv:1711.11581*.

Ma, T., J. Shi, and D. Steurer (2016): "Polynomial-time Tensor Decompositions with Sum-of-Squares," arXiv:1610.01980, abstract appeared in FOCS 2016.

Raghavendra, P. and M. Yau (2020): "List decodable learning via sum of squares," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 161–180.