Estimation Contracts for Outlier-Robust Geometric Perception

Luca Carlone*

December 30, 2024

Abstract

Outlier-robust estimation is a fundamental problem and has been extensively investigated by statisticians and practitioners. The last few years have seen a convergence across research fields towards "algorithmic robust statistics", which focuses on developing tractable outlier-robust techniques for high-dimensional estimation problems. Despite this convergence, research efforts across fields have been mostly disconnected from one another. This monograph bridges recent work on certifiable outlier-robust estimation for geometric perception in robotics and computer vision with parallel work in robust statistics. In particular, we adapt and extend recent results on *robust linear regression* (applicable to the low-outlier regime with $\ll 50\%$ outliers) and *list-decodable regression* (applicable to the high-outlier regime with $\gg 50\%$ outliers) to the setup commonly found in robotics and vision, where (i) variables (*e.g.,* rotations, poses) belong to a non-convex domain, (ii) measurements are vector-valued, and (iii) the number of outliers is not known a priori. The emphasis here is on performance guarantees: rather than proposing radically new algorithms, we provide conditions on the input measurements under which modern estimation algorithms (possibly after small modifications) are guaranteed to recover an estimate close to the ground truth in the presence of outliers. These conditions are what we call an "estimation contract". The monograph also provides numerical experiments to shed light on the applicability of the theoretical results and to showcase the potential of list-decodable regression algorithms in geometric perception. Besides the proposed extensions of existing results, we believe the main contributions of this monograph are (i) to unify parallel research lines by pointing out commonalities and differences, (ii) to introduce advanced material (*e.g.,* sum-of-squares proofs) in an accessible and self-contained presentation for the practitioner, and (iii) to point out a few immediate opportunities and open questions in outlier-robust geometric perception.

Contents

1 Introduction

Geometric perception is the problem of estimating unknown geometric models (*e.g.,* poses, rotations, 3D structure) from sensor data (*e.g.,* camera images, lidar scans, inertial data, wheel odometry). Geometric perception has been at the center stage of robotics and computer vision research since their inception, and includes problems such as object pose (and possibly shape) estimation [1, 2], robot or camera motion estimation [3], sensor calibration [4], Simultaneous Localization And Mapping (SLAM) [5], and Structure from Motion (SfM) [6], to mention a few.

At its core, geometric perception solves an estimation problem, where, given measurements $\boldsymbol{y}_i, i = 1, \ldots, n$, one has to compute a variable of interest $\boldsymbol{x}^\circ$ (the "ground truth"). For instance, in an object pose estimation problem, $\boldsymbol{x}^\circ$ is the to-be-computed 3D pose of the object (say, a car), while the $\boldsymbol{y}_i$'s might be observations of relevant points on the object (*e.g.,* the wheels and the headlights of the car). The unknown $\boldsymbol{x}^\circ$ and the measurements $\boldsymbol{y}_i$ are related by a *measurement (or generative) model.* In this monograph, we focus our attention on the common case where the measurements are vector-valued, *i.e.,* $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$, and the noise is additive, leading to measurement models in the form:

$$\boldsymbol{y}_i = f_i(\boldsymbol{x}^\circ) + \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{y}_i \in \mathbb{R}^{d_y} \quad \text{and} \quad \boldsymbol{x}^\circ \in \mathbb{X} \subseteq \mathbb{R}^{d_x}, \tag{1}$$

where $f_i(\cdot)$ is a known function, $\boldsymbol{\epsilon}$ is the measurement noise, and $\mathbb{X}$ is the domain of $\boldsymbol{x}^\circ$ (*e.g.,* the set of 3D poses in an object pose estimation problem). As we will see in Section 3, many geometric perception problems have measurement models in the form of eq. (1).[1]

When the noise in (1) is zero-mean and Gaussian, the maximum likelihood estimate of $\boldsymbol{x}^\circ$ can be computed via standard least squares:[2]

$$\boldsymbol{x}_{\mathsf{LS}} = \underset{\boldsymbol{x} \in \mathbb{X}}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{y}_i - f_i(\boldsymbol{x})\|_2^2 . \tag{LS}$$

While problem (LS) can be still hard to solve (*e.g.,* due to potential non-convexity of $f_i(\cdot)$ or $\mathbb{X}$), its structure —at least for common geometric perception problems— has been extensively studied in robotics and vision, and the literature offers a broad range of solvers, including closed-form solutions [8], iterative local solvers [9], minimal solvers [10], and convex relaxations [7, 11, 2, 12, 13, 14].

Outlier-robust estimation. In practice, many of the measurements fed to the estimation process are *outliers, i.e.,* they largely deviate from the measurement model (1) and possibly do not carry any information about $\boldsymbol{x}^\circ$. In robotics and vision, the measurements $\boldsymbol{y}_i$ are the result of a pre-processing of the raw sensor data; such preprocessing is often referred to as the *perception front-end,* while the estimation algorithms that compute $\boldsymbol{x}^\circ$ from the $\boldsymbol{y}_i$'s are referred to as the *perception back-end.* For instance, in an object pose estimation problem, the perception front-end extracts the position of relevant features $\boldsymbol{y}_i$ on the object from raw image pixels (typically using a neural network), while the back-end computes the object pose given the $\boldsymbol{y}_i$'s. The perception front-end is prone to errors (*e.g.,* the network may mis-detect the wheels of the car in the image), resulting in measurements $\boldsymbol{y}_i$ with large errors. In the presence of outliers, the least squares estimator (LS) is

---

[1]Assuming additive noise comes at a small loss of generality, *e.g.,* in SLAM and rotation averaging the measurements belong to a smooth manifold rather than a vector space and the noise is multiplicative. However, even in these cases, the resulting outlier-free formulations —under suitable noise assumptions— lead to standard least squares [7, 4], hence we believe adapting the results in this monograph to those setups is indeed possible, see Section 9.

[2]Without loss of generality, we assume $\boldsymbol{\epsilon}$ to have an isotropic Gaussian distribution with identity covariance, but arbitrary covariances can be easily accommodated by rescaling $\boldsymbol{y}_i$ and $f_i(\cdot)$ by the square root of the inverse covariance.

known to produce grossly incorrect results, hence it is desirable to adopt an outlier-robust estimator that can correctly estimate $\boldsymbol{x}^{\circ}$ in the presence of many outliers. In this monograph, we do not make assumptions on the nature of the outliers and consider the worst case where a fraction β of measurements is arbitrarily corrupted, a setup commonly referred to as the *strong adversary model* in statistics and learning [15].

The robust statistics lens. Classical robust statistics [16, 17, 18, 19] provides many alternative formulations to (LS) that allow regaining robustness to outliers. For instance, if the number of outliers is known, say a fraction β of the n measurements is corrupted, we can use the *Least Trimmed Squares* (LTS) estimator [18] to compute an outlier-robust estimate:

$$\boldsymbol{x}_{\mathsf{LTS}} = \arg\min_{\substack{\boldsymbol{\omega} \in \{0;1\}^n \\ \boldsymbol{x} \in \mathbb{X}}} \ \sum_{i=1}^{n} \omega_i \cdot \|\boldsymbol{y}_i - f_i(\boldsymbol{x})\|_2^2 \ , \quad \text{subject to} \quad \sum_{i=1}^{n} \omega_i = \alpha\,n \ , \tag{LTS}$$

where we defined the *inlier rate* $\alpha \triangleq 1 - \beta$, and introduced binary variables $\boldsymbol{\omega} \in \{0;1\}^n$ which are in charge of selecting the best $\alpha\,n$ measurements (when $\omega_i = 1$, the i-th measurement is selected as an inlier by (LTS), while $\omega_i = 0$ otherwise); in words, (LTS) selects the $\alpha\,n$ measurements that induce the smallest error for some estimate $\boldsymbol{x}$ and disregards the remaining measurements as outliers. Unfortunately, the optimization problem (LTS), as well as many other popular outlier-robust formulations, are NP-hard [20] and for a long while no tractable algorithm was available for high-dimensional outlier-robust estimation problems (*e.g.,* in the problems we discuss in Section 3 and Section 9, $\boldsymbol{x}$'s dimension ranges from 9 to potentially more than a thousand). In recent years, *algorithmic robust statistics* came to the rescue, by proposing polynomial-time algorithms for outlier-robust estimation with strong performance guarantees, including [15, 21, 22, 23, 24]. For instance, while not explicitly recognized in the paper, the algorithm by Klivans *et al.* [15] can be understood as a convex relaxation for problem (LTS) for the case where $f_i(\cdot)$ is a real-valued linear function. Many of these works use Lasserre's moment relaxation [25] as an algorithmic workhorse, and adopt the dual view of sum-of-squares relaxations [26] to prove bounds on the quality of the estimates.

The computer vision lens. In typical robotics and vision applications, the number of outliers is unknown, therefore outlier-robust estimators have to simultaneously look for a suitable estimate of $\boldsymbol{x}^{\circ}$ while searching for a large set of inliers. In computer vision, a common formulation for outlier-robust estimation with unknown number of outliers is *consensus maximization* [27], which searches for the largest set of inliers such that the measurements selected as inliers have a low error with respect to some estimate:

$$\boldsymbol{x}_{\mathsf{MC}} = \arg\max_{\substack{\boldsymbol{\omega} \in \{0;1\}^n \\ \boldsymbol{x} \in \mathbb{X}}} \ \sum_{i=1}^{n} \omega_i \ , \quad \text{subject to} \quad \omega_i \cdot \|\boldsymbol{y}_i - f_i(\boldsymbol{x})\|_2^2 \leq \bar{c}^2 \ , \tag{MC}$$

where the given constant $\bar{c} \geq 0$ is the maximum error for a measurement to be considered an inlier. Problem (MC) has been shown to be inapproximable [28, 29], and the literature has been traditionally split between fast heuristics (which do not provide performance guarantees) and globally optimal solvers (which can compute optimal solutions but run in worst-case exponential time). The recent work [30] shows that for common geometric perception problems, (MC) can be written as a polynomial optimization problem (POP) and relaxed via Lasserre's moment relaxation. The key insight behind [30], reviewed in Section 3, is that —for common perception problems— the domain $\mathbb{X}$ is a basic semi-algebraic set (*i.e.,* it can be written as a set of polynomial inequalities), while with a suitable parametrization, the function $f_i(\cdot)$ becomes a (vector-valued) linear function.

The robotics lens. In robotics, the go-to approach for outlier-robust estimation has been the use of M-estimators [17], which replace the least squares cost in (LS) with a robust loss function. In this monograph we focus on a particular choice of robust loss function, the *truncated least squares (or truncated quadratic)* cost:

$$
\begin{aligned}
\boldsymbol{x}_{\mathsf{TLS}} &= \arg\min_{\boldsymbol{x}\in\mathbb{X}} \sum_{i=1}^{n} \min\left(\|\boldsymbol{y}_i - f_i(\boldsymbol{x})\|_2^2 \ , \ \bar{c}^2 \right) \\
&= \arg\min_{\substack{\boldsymbol{\omega}\in\{0;1\}^n \\ \boldsymbol{x}\in\mathbb{X}}} \sum_{i=1}^{n} \omega_i \cdot \|\boldsymbol{y}_i - f_i(\boldsymbol{x})\|_2^2 + (1-\omega_i)\cdot \bar{c}^2 \ ,
\end{aligned}
\tag{TLS}
$$

where the objective is the pointwise minimum of a quadratic and a constant function, *i.e.,* it is quadratic for small *residuals* $\|\boldsymbol{y}_i - f_i(\boldsymbol{x})\|_2 \leq \bar{c}$, and becomes constant for large residuals. In the second line in (TLS) we noticed that the truncated least squares cost can be equivalently rewritten using auxiliary binary variables $\boldsymbol{\omega}$, by observing that for two numbers a, b, $\min(a,b) = \min_{\omega\in\{0;1\}} \omega \cdot a + (1-\omega)\cdot b$. Also problem (TLS) has been shown to be inapproximable in the worst case [28]. While traditionally problem (TLS) has been attacked using local solvers [31] or continuation schemes [32], recent work [30, 33, 1, 34, 35] has shown that for common perception problems, (TLS) can be written as a POP and relaxed via Lasserre's moment relaxation. More surprisingly, many works have empirically observed the relaxation to be tight [30, 33, 34], at least for reasonable levels of noise and outliers, with very recent work [36] providing initial theoretical results to support such empirical evidence, at least for the specific problem of rotation search. However, the performance of these estimators is commonly demonstrated via empirical evaluation, and the literature is still lacking more general theoretical guarantees on the quality of the resulting estimates.

Catalyst, convergence, and contribution. Despite the heterogeneity of the formulations reviewed above, we observe that recent years have witnessed a convergence across fields towards designing tractable algorithms for high-dimensional outlier-robust estimation using moment relaxations. A few examples include [15, 37, 30, 33, 1, 34, 35, 36, 38]. Such a convergence has been triggered by the progress in polynomial optimization via moment and sum-of-squares relaxations, starting from the seminal works [25, 26, 39, 40, 41]. At the same time, research across fields has remained disconnected, with researchers being mostly unaware of the parallel work in other areas.

The goal of this monograph is to bridge this gap and connect geometric perception problems in robotics and vision to novel tools in outlier-robust statistics. Towards this goal we adapt and extend recent results from robust statistics to the setup and formulations commonly found in robotics and vision. For the case with low outlier rates (*i.e.,* $\beta \ll 0.5$), we adapt results from [15], which considers outlier-robust regression using least trimmed squares (LTS) with scalar linear measurements, to the robotics setup where the measurements are vector valued and the variables belong to a non-convex domain; we also develop a simple bound on the distance of the estimate from the ground truth (while [15] focuses on bounding the residual errors for the inliers). Then, we extend these results to the case where the number of outliers is unknown. In particular, we compute bounds on the estimation error (*i.e.,* the distance between the estimate and $\boldsymbol{x}^{\circ}$) for (MC) and (TLS). These results constitute the first general performance guarantees for the convex relaxations [30, 33, 34], going beyond the empirical observations in [30] and the problem-specific optimality guarantees in [34, 36].

Then we consider the case with high outlier rates (*i.e.,* $\beta \gg 0.5$), where a majority of the measurements are outliers. While in robotics and vision it has been observed that with random (*i.e.,* non-adversarial) outliers, the point estimators (MC) and (TLS) are still able to retrieve

good estimates for $\boldsymbol{x}^{\circ}$ [30, 33, 34, 1], in the presence of adversarial outliers, the estimate resulting from (LTS), (MC), and (TLS) can be arbitrarily far from the ground truth: intuitively, since the outliers constitute the majority of the measurements, they can agree on an arbitrary $\boldsymbol{x}$ and form a large set of mutually consistent measurements that are picked as solution to (LTS), (MC), and (TLS).[3] In robotics and related fields, this setup has been recognized to require computing multiple estimates, in order to find one that is close to the ground truth, ranging from early work on multi-hypothesis target tracking [42] and particle filters [43], to recent work on multi-hypothesis smoothing [44, 45]. However, none of these works simultaneously provide tractable algorithms and performance guarantees for the resulting estimates. In this monograph, we connect to the recent literature on *list-decodable regression* [37], which proposes polynomial-time estimators that return a small list of estimates such that with high probability at least one of the estimates is close to the ground truth. In particular, we provide an adaptation of the results in [37] to account for vector-valued measurements.

Finally, we present numerical experiments on a canonical geometric perception problem to shed light on the theoretical results. The experiments provide encouraging evidence that many of the assumptions supporting the theoretical analysis (*e.g.,* certifiable hypercontractivity) are often satisfied by real data. At the same time, they reveal a large gap between theory (which mostly guarantees performance for high-order, computationally expensive moment relaxations) and practice (where low-order relaxations already exhibit impressive performance). Our numerical evaluation also provides the first empirical evidence that a sparse low-order moment relaxation for list-decodable regression (based on a modified version of the algorithm proposed in [37]) is able to accurately recover estimates in geometric perception problems with high outlier rates, where (LTS), (MC), and (TLS) are doomed to fail. Moreover, the experiments show that if the measurements are generated by multiple estimates (*e.g.,* different subsets of measurements are generated by different variables $\boldsymbol{x}^{\circ}$), then our sparse moment relaxation for list-decodable regression is able to simultaneously recover *all* the estimates generating the data.[4] We release open-source code to reproduce our numerical experiments, including an implementation of key algorithms covered in this monograph at https://github.com/MIT-SPARK/estimation-contracts.

We remark that the emphasis in this monograph is on performance guarantees. We do not present new algorithms (we mostly propose small modifications to existing algorithms) but rather try to address the question: under which conditions on the input measurements can we guarantee that modern outlier-robust estimation algorithms based on moment relaxations recover an estimate close to the ground truth in the presence of outliers? These conditions are what we call an "estimation contract". Besides the proposed extensions of existing results, we believe the main contributions of this monograph are (i) to unify parallel research lines by pointing out commonalities and differences, (ii) to introduce advanced material (*e.g.,* sum-of-squares proofs) in an accessible and self-contained presentation for the practitioner, and (iii) to point out a few immediate opportunities and open questions in outlier-robust geometric perception. This "unification" is expected to benefit both practitioners and researchers in robust statistics. On the robotics and computer vision side, this monograph provides new and fairly general performance guarantees for robust estimation algorithms

---

[3]Note that the case with a high number of adversarial outliers is often the one encountered in practice in robotics and vision: think about a motion estimation problem where the robot has to estimate its motion from point features detected by the camera [3]: if there is a large moving object in front of the camera, most features may fall on the moving object (rather that on the static portion of the scene), leading to incorrect motion estimates.

[4]With reference to the motion estimation example in footnote 3, such an algorithm would simultaneously recover the motion of all the objects in the scene, rather than just the motion with respect to the object capturing most point features, which would be quite useful in practical applications.

based on moment relaxations, applied to geometric perception problems. Moreover, the monograph reviews a new proof system (based on sum-of-squares proofs) that provides a richer language to discuss properties of moment relaxations beyond the typical analysis based on a manual design of dual certificates [34, 36, 46, 7]. Furthermore, it positions list-decodable regression based on moment relaxations as a useful and computationally tractable tool for multi-hypotheses estimation. On the robust statistics side, we hope the reader will be intrigued by the remarks about the practical performance and the empirical tightness of the moment relaxation of (TLS) (discussed in greater detail in [30]) and the practical performance of a low-order relaxation for list-decodable regression, which we believe deserve further investigation. We also hope to attract further attention towards the case where the number of outlier is unknown and the variables are confined to semi-algebraic sets, which is the setup commonly encountered in robotics and vision problems.

Monograph structure. Section 2 starts by reviewing related works across fields. Section 3 showcases the fact that many estimation problems in robotics and vision can be formulated using a linear measurement model with variables belonging to a basic semi-algebraic set. Section 4 introduces notation and preliminaries (while postponing as many details as possible to the appendix). Section 5 succinctly states the problem of outlier-robust estimation and our quest for estimation contracts. Section 6 studies the case with low outlier rates and provides error bounds for (LTS), (MC), and (TLS). Section 7 studies the case with high outlier rates and adapts results from list-decodable regression. Section 8 presents numerical experiments on a rotation search problem. Section 9 discusses opportunities and open problems, and Section 10 concludes the monograph. The appendices provide background information regarding moment relaxations (Appendix A), pseudo-distributions (Appendix B), and sum-of-squares proofs (Appendix C), and contain technical proofs (Appendices D to I) and algorithmic details about sparse list-decodable estimation (Appendix J).

Learning paths. The main body of this monograph (up to Section 10) is designed to be self-contained, and readers interested in getting a bird's-eye view of the technical tools underlying our results, the corresponding performance guarantees, and sparse list-decodable estimation can focus their attention on these sections. Expert readers from statistics might find the numerical experiments of particular interest. Junior researchers in computer vision and robotics interested in advancing this research are recommended to carefully read Appendices A to C, which should serve as a gentle introduction to the key technical tools. Beyond this monograph, the book [47] and monograph [9] provide necessary background information about geometric perception, the book [48] provides an in-depth introduction to semidefinite programming and sum-of-squares relaxations, and the notes [49] and monograph [50] provide a more formal introduction to sum-of-squares proofs.

2 Related Work

Outlier-robust estimation in robotics and computer vision. Traditional algorithms for outlier-robust estimation for geometric perception can be divided into *fast heuristics* and *globally optimal solvers*. Two general frameworks for designing fast heuristics are RANSAC [51] and *graduated non-convexity* (GNC) [52, 32, 28]. RANSAC solves problem (MC) by repeatedly sampling a minimal set of measurements, computing an estimate x from the minimal set, and searching for a large set of measurements that agrees with the estimate x [27]; while being a well-established algorithm, RANSAC mostly applies to low-dimensional problems, since the expected number of iterations required by RANSAC to find an outlier-free set of measurements grows exponentially in the size of the minimal measurement set; moreover, the number of RANSAC iterations also grows exponentially with the outlier rate β [53]. GNC solves M-estimation problems, including (TLS), via a continuation

scheme that starts from a convex approximation of the cost function and then gradually recovers the robust cost. GNC can scale to high-dimensional problems and applies to a broad class of loss functions —including adaptive loss functions [54, 55]. However, the effectiveness of the continuation scheme is problem-dependent. Iterative local optimization is also a popular fast heuristics for the case where an initial guess is available [56, 57, 58, 59]. Approximate but deterministic algorithms have also been designed to solve consensus maximization [60]. On the other hand, globally optimal solvers are guaranteed to retrieve optimal solutions for outlier-robust estimation but run in worst-case exponential time. These solvers are typically designed using Branch and Bound or mixed-integer programming [61, 62, 63, 64, 65, 66, 67, 68]. Enqvist *et al.* [69] present an outlier-robust estimator that runs in polynomial time with respect to the number of measurements, but still has exponential complexity in the dimension of the variable to be estimated (see also [70, 71]). Chin *et al.* [72] frame (MC) as a tree search problem and propose an approach based on A^* search.

Certifiably optimal outlier-robust algorithms [30, 33, 1, 34, 35] have recently emerged as a way to obtain optimal solutions to certain outlier-robust problems in polynomial time (both with respect to the number of measurements and the dimension of the variable to be estimated). These approaches relax non-convex outlier-robust formulations, including (MC) and (TLS), into a convex optimization. The key insight behind these algorithms is twofold: (i) for several estimation problems arising in robotics and vision, we can rewrite the optimization problems arising in outlier-robust estimation as *polynomial optimization problems* (POP), and (ii) for these POPs, Lasserre's moment relaxation [25] is empirically seen to be tight at the lowest relaxation order (*i.e.,* order 2), hence providing a tractable way to compute robust estimates. These works mostly provide a posteriori optimality certificates [30, 33] (*i.e.,* they solve the relaxation, compute a rounded estimate, and then compute a suboptimality gap for that specific estimate, possibly certifying its optimality), while the papers [34, 36] provide problem-specific a priori conditions under which the relaxation is tight.

Outlier-robust estimation in robust statistics. Outlier-robust estimation has been studied in robust statistics, starting from the seminal work of Huber, Tukey, and Rousseeuw [16, 17, 18, 19], among many others. However, these classical frameworks do not immediately lead to tractable algorithms and are often provably hard to solve [20, 28, 27], leading to algorithms whose complexity increases exponentially in the dimension of $\boldsymbol{x}$. The growing interest towards high-dimensional outlier-robust estimation has triggered a large number of recent works on "algorithmic robust statistics", which focus on the design of tractable algorithms for high-dimensional estimation with outliers. Early algorithms along this line have focused on clustering and moment estimation (*e.g.,* how to robustly estimate mean and covariance of a distribution given samples) [73, 74, 75, 76, 77, 78, 79] and subspace learning for classification in the presence of malicious noise [80, 81, 82].

More relevant to this monograph is the recent work on *robust linear regression* [15, 21, 22, 23, 24, 83], where one has to compute an estimate of $\boldsymbol{x}^\circ$ given an outlier-corrupted set of linear measurements:

$$y_i = \boldsymbol{a}_i^\mathsf{T} \boldsymbol{x}^\circ + \epsilon, \quad \text{with} \quad y_i \in \mathbb{R} \quad \text{and} \quad \boldsymbol{x}^\circ \in \mathbb{R}^{d_x}, \tag{2}$$

where y_i are given scalar measurements, $\boldsymbol{a}_i$ are given vectors of suitable dimension, and ϵ is the measurement noise. Typical contamination models studied in the literature include the *Huber contamination model* (*e.g.,* [84]), which assumes that a fraction of measurements are randomly generated by an unknown outlier distribution, and the *strong adversary model*, in which the outlier-generation mechanism has access to the inliers and can replace a given fraction of them with arbitrary outliers. The literature on outlier-robust linear regression in the low-outlier regime ($\beta \ll 0.5$)

includes approaches based on iterative outlier filtering [21, 23], robust gradient estimation [22], hard thresholding [24, 85, 86], ℓ_1-regression [87, 83, 88], and moment/sum-of-squares relaxations [15]. Our interest towards moment/sum-of-squares relaxations is motivated by that fact that the framework in [15] can accommodate a large class of linear measurement models —more precisely, fairly general distributions for the vectors $\boldsymbol{a}_i$ in (2)— while related work often takes stronger assumptions on $\boldsymbol{a}_i$ (*e.g.,* [23] assumes $\boldsymbol{a}_i$'s are sampled from a Gaussian distribution); moreover, the corresponding algorithms operate in the strong adversary model, which is useful to derive worst-case guarantees for geometric perception. In the high-outlier regime ($\beta \gg 0.5$), it is still possible to recover a good estimate when the outliers are not adversarial. The works [87, 88] provide approaches based on ℓ_1-minimization to recover good estimates assuming oblivious outliers (rather than adversarial), and under relatively strong assumptions on the measurements ([88] assumes sub-orthogonality, [87] assumes the vectors $\boldsymbol{a}_i$ to be drawn from an isotropic Gaussian). The works [89, 90, 91] also assume oblivious outliers and analyze the Huber loss estimator ([90] assumes Gaussian $\boldsymbol{a}_i$, [89] requires the column span of the matrix staking all the vectors $\boldsymbol{a}_i$ not to contain approximately sparse vectors, [91] assumes a bounded or sub-Gaussian distribution for the $\boldsymbol{a}_i$'s). The works [37, 92] are the first to provide polynomial-time algorithms for *list-decodable* linear regression with adversarial outliers, where the estimator returns multiple hypotheses such that at least one of the hypotheses is close to the ground truth $\boldsymbol{x}^\circ$. List-decodable learning was originally introduced in [93] and was also studied in the context of moment estimation in [76, 94, 77].

3 Motivating Problems

This section shows that many foundational problems in geometric perception can be formulated as *linear* estimation problems with variables belonging to a *basic semi-algebraic set* (*i.e.,* a set that can be described by a finite number of polynomial inequality constraints).[5] Mathematically, we will formulate the measurement model for many problems of interest as:

$$\boldsymbol{y}_i = \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ + \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{y}_i \in \mathbb{R}^{d_y} \quad \text{and} \quad \boldsymbol{x}^\circ \in \mathbb{X} \subseteq \mathbb{R}^{d_x}. \tag{3}$$

This observation will be crucial towards adapting existing results in robust statistics to the geometric perception setup. Indeed, the measurement model (3) is essentially the same of the one used in robust regression, see eq. (2), with the exception that measurements $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$ are vector-valued, and the variable $\boldsymbol{x}^\circ$ belongs to a specific domain $\mathbb{X}$ (typically, a smooth manifold) rather than $\mathbb{R}^{d_x}$. Towards recasting many estimation problems in geometric perception as in eq. (3), we start by restating the well-known fact that —for common variables of interest in robotics and vision— the domain $\mathbb{X}$ is a basic semi-algebraic set (see *e.g.,* [30, 95, 13, 96, 97]).

Fact 1 (Variables in geometric perception)**.** *The d-dimensional* Special Orthogonal group $\mathrm{SO}(d) \triangleq \{\boldsymbol{R} \in \mathbb{R}^{d\times d} \mid \boldsymbol{R}^\mathsf{T}\boldsymbol{R} = \mathbf{I}_d, \det(\boldsymbol{R}) = +1\}$ *(*i.e.*, the group of rotations), and the* Special Euclidean group $\mathrm{SE}(d) \triangleq \left\{ \left[\begin{smallmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{smallmatrix} \right] \in \mathbb{R}^{(d+1)\times(d+1)} \mid \boldsymbol{R} \in \mathrm{SO}(d), \boldsymbol{t} \in \mathbb{R}^d \right\}$ *(*i.e.*, the group of poses and rigid transformations) are basic semi-algebraic sets.[6] Moreover, several geometric constraints (*e.g.*, field-of-view or maximum distance constraints) can be written as basic semi-algebraic sets.*

---

[5]This observation constitutes the basis for many certifiable solvers for outlier-free and outlier-robust estimation that have been developed in robotics and vision, see [30, 7] and the references therein.

[6]More precisely, $\mathrm{SO}(d)$ and $\mathrm{SE}(d)$ are *algebraic varieties*, *i.e.,* sets that can be described by a finite set of polynomial *equality* constraints $\{f_1(\boldsymbol{x}) = 0, f_2(\boldsymbol{x}) = 0, \ldots, f_m(\boldsymbol{x}) = 0\}$. Note that we can rewrite a variety as a basic semi-algebraic set by replacing each equality constraint $f_i(\boldsymbol{x}) = 0$ with two inequality constraint $f_i(\boldsymbol{x}) \geq 0$ and $f_i(\boldsymbol{x}) \leq 0$, hence a variety can be understood as a special case of a basic semi-algebraic set.

Now we observe that several geometric perception problems can be written as *linear* models, akin to eq. (3) (we will discuss few more examples, including SLAM and rotation averaging in Section 9). The reader familiar with geometric perception can safely skip this section.

Example 1 (Rotation search (a.k.a. the Wahba problem))**.** *Estimate the rotation $\boldsymbol{R} \in \mathrm{SO}(3)$ that aligns pairs of 3D points $(\boldsymbol{a}_i, \boldsymbol{b}_i)$, $i = 1, \ldots, n$. The measurement model for the (inlier) measurements is given by:*

$$\boldsymbol{b}_i = \boldsymbol{R}\boldsymbol{a}_i + \boldsymbol{\epsilon} \quad \xRightarrow{\boldsymbol{x}^\circ \triangleq vec(\boldsymbol{R})} \quad \overbrace{\boldsymbol{b}_i = (\boldsymbol{a}_i^\mathsf{T} \otimes \mathbf{I}_3)\boldsymbol{x}^\circ + \boldsymbol{\epsilon},}^{\text{same form as (3)}} \tag{4}$$

where $\boldsymbol{\epsilon}$ is the measurement noise. In (4), we used the vectorization operator $vec(\cdot)$ to transform a 3D matrix into a vector and manipulated the expression using standard vectorization properties. Rotation search arises, for instance, in satellite attitude estimation [98] and image stitching [34].

Example 2 (3D point cloud registration)**.** *Estimate the rigid transformation $(\boldsymbol{R}, \boldsymbol{t})$, with $\boldsymbol{R} \in \mathrm{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$, that aligns pairs of 3D points $(\boldsymbol{a}_i, \boldsymbol{b}_i)$, $i = 1, \ldots, n$. The measurement model for the (inlier) measurements is:*

$$\boldsymbol{b}_i = \boldsymbol{R}\boldsymbol{a}_i + \boldsymbol{t} + \boldsymbol{\epsilon} \quad \xRightarrow{\boldsymbol{x}^\circ \triangleq \left[\begin{smallmatrix} vec(\boldsymbol{R}) \\ \boldsymbol{t} \end{smallmatrix}\right]} \quad \overbrace{\boldsymbol{b}_i = \left[\begin{array}{cc} \boldsymbol{a}_i^\mathsf{T} \otimes \mathbf{I}_3 & \mathbf{I}_3 \end{array}\right] \boldsymbol{x}^\circ + \boldsymbol{\epsilon},}^{\text{same form as (3)}} \tag{5}$$

Point-to-plane 3D registration can be similarly formulated using a linear model involving a rigid transformation [30]. Registration problems are commonly encountered in instance-level object pose estimation, scan-matching for 3D reconstruction, and (stereo or RGB-D) visual odometry [1].

Example 3 (3D-3D category-level object pose and shape estimation)**.** *Estimate the rigid transformation $(\boldsymbol{R}, \boldsymbol{t})$, with $\boldsymbol{R} \in \mathrm{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$, and the shape parameters $\boldsymbol{c} \in \mathbb{R}^K$ (describing the shape of a 3D object) from 3D point measurements $\boldsymbol{b}_i$, $i = 1, \ldots, n$. The generative model for the (inlier) measurements is:*

$$\boldsymbol{b}_i = \boldsymbol{R}\boldsymbol{S}_i\boldsymbol{c} + \boldsymbol{t} + \boldsymbol{\epsilon} \quad \xRightarrow{\boldsymbol{t}' \triangleq \boldsymbol{R}^\mathsf{T}\boldsymbol{t}\,,\, \boldsymbol{\epsilon}' \triangleq \boldsymbol{R}^\mathsf{T}\boldsymbol{\epsilon}\,,\, \boldsymbol{x}^\circ \triangleq \left[\begin{smallmatrix} vec(\boldsymbol{R}^\mathsf{T}) \\ \boldsymbol{t}' \\ \boldsymbol{c} \end{smallmatrix}\right]} \quad \mathbf{0} = \left[\begin{array}{ccc} -\boldsymbol{b}_i^\mathsf{T} \otimes \mathbf{I}_3 & \mathbf{I}_3 & \boldsymbol{S}_i \end{array}\right] \boldsymbol{x}^\circ + \boldsymbol{\epsilon}', \tag{6}$$

where $\boldsymbol{S}_i \in \mathbb{R}^{3 \times K}$ is a matrix of given basis shapes (such that the final shape $\boldsymbol{S}_i\boldsymbol{c}$, $i = 1, \ldots, n$, is written as a linear combination of the basis shapes). Note that for (6) to fall in the class of problems (3), we have to assume that $\boldsymbol{\epsilon}$ has an isotropic distribution, such that the distribution of $\boldsymbol{\epsilon}' \triangleq \boldsymbol{R}^\mathsf{T}\boldsymbol{\epsilon}$ does not depend on the unknown $\boldsymbol{R}$. The model in (6) is known as the active shape model *[99], and finds application in face detection, human pose estimation, and object pose and shape estimation, see [2, 97, 11, 100] among others. In these problems, it is not uncommon for the number of shapes K to be large, e.g., $K \gg 100$.*

Example 4 (Absolute pose estimation)**.** *Estimate the camera pose $(\boldsymbol{R}, \boldsymbol{t})$, with $\boldsymbol{R} \in \mathrm{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$, from (calibrated) pixel observations, written as unit-norm vectors $\boldsymbol{u}_i$, picturing known 3D points $\boldsymbol{a}_i \in \mathbb{R}^3$, $i = 1, \ldots, n$. The generative model for the (inlier) measurements is:*

$$\lambda_i \boldsymbol{u}_i = \boldsymbol{R}\boldsymbol{a}_i + \boldsymbol{t} + \boldsymbol{\epsilon} \quad \xRightarrow{\boldsymbol{\epsilon}_i' \triangleq [\boldsymbol{u}_i]_\times \boldsymbol{\epsilon}\,,\, \boldsymbol{x}^\circ \triangleq \left[\begin{smallmatrix} vec(\boldsymbol{R}) \\ \boldsymbol{t} \end{smallmatrix}\right]} \quad \overbrace{\mathbf{0} = \left[\begin{array}{cc} [\boldsymbol{u}_i]_\times (\boldsymbol{a}_i^\mathsf{T} \otimes \mathbf{I}_3) & [\boldsymbol{u}_i]_\times \end{array}\right] \boldsymbol{x} + \boldsymbol{\epsilon}_i',}^{\text{same form as (3)}} \tag{7}$$

Intuitively, the measurement model on the left describes the fact that —up to an unknown scale λ_i— the pixel measurement $\boldsymbol{u}_i$ pictures the 3D point $\boldsymbol{a}_i$ after it is transformed to the camera frame according to the camera pose $(\boldsymbol{R}, \boldsymbol{t})$. While the model on the left is already linear, on the right-hand side we algebraically eliminated the scale factors λ_i and obtained a lower-dimensional linear measurement model by multiplying both sides by the orthogonal projector $[\boldsymbol{u}_i]_\times \triangleq \mathbf{I}_3 - \boldsymbol{u}_i \boldsymbol{u}_i^\mathsf{T}$, which is such that $[\boldsymbol{u}_i]_\times \boldsymbol{u}_i = \mathbf{0}$. The absolute pose estimation problem arises in camera localization in known scenes and object pose estimation from camera images, see e.g., [101, 102, 103].

Two remarks are in order. First, we observe that while the adoption of the linear model in (3) with variables in $\mathbb{X}$ might seem an obvious choice, this has not necessarily been the go-to approach in robotics and vision. In many cases, researchers still prefer non-linear measurement models over Euclidean space,[7] rather than a linear model over the semi-algebraic domain $\mathbb{X}$, since the former leads to *unconstrained* nonlinear least squares problems that can be quickly solved by local solvers when an initial guess is available [9]. The second remark is that —in the outlier-free case— all the geometric perception examples in this section can be considered solved; in particular, rotation search and 3D registration admit a closed-form solution [8]; absolute pose estimation can be solved globally using Gröbner basis [101]; 3D-3D category-level perception can be solved via a tight convex relaxation [100]. On the other hand, some of these problems remain challenging in the presence of outliers and are still the subject of active research, see [53, 1, 100] and the references therein.

We conclude this section with an assumption that is required for the theoretical analysis and practical performance of the machinery used in this monograph, *i.e.,* moment/sum-of-squares relaxations.

Assumption 2 (Explicitly bounded domain)**.** *In this monograph we assume that the domain $\mathbb{X}$ is* explicitly bounded *(or* Archimedian*), meaning that it contains a constraint in the form $\|\boldsymbol{x}\|_2^2 \leq M_x^2$ for some finite constant $M_x > 0$.*

This assumption is typically not restrictive since many geometric variables already belong to bounded sets. For instance, rotations $\boldsymbol{R} \in \mathrm{SO}(3)$ satisfy $\|vec(\boldsymbol{R})\|_2^2 = 3$; the shape vector in Example 3 is typically assumed to belong to the probability simplex, hence it satisfies $\|\boldsymbol{c}\|_2^2 \leq 1$; finally, translations can be assumed to be bounded since the sensors producing the measurements in the examples above have finite range.

4 Preliminaries on Moment Relaxations and Sum-of-Squares Proofs

This section reviews the three main ingredients of modern techniques for outlier-robust estimation: semidefinite programming, moment relaxations, and sum-of-squares proofs. In the next sections, we will use these concepts to state outlier-robust estimation algorithms (using moment relaxations, which lead to relaxing our estimation problems to tractable semidefinite programs) and to analyze their performance (using sum-of-squares proofs). We keep this presentation short and pragmatic, and refer the interested reader to the appendix and to specialized references [104, 41, 49] for details.

Notation. We use lowercase characters (*e.g.,* y) to denote real scalars, bold lowercase characters (*e.g.,* $\boldsymbol{y}$) for real (column) vectors, and bold uppercase characters (*e.g.,* $\boldsymbol{A}$) for real matrices. $\mathbf{I}_d$ denotes the identity matrix of size $d \times d$, $\mathbf{1}_d$ denotes the vector of ones of size d, $\mathbf{0}$ denotes the all-zero vector or matrix of appropriate size. The symbol $\otimes$ denotes the Kronecker product. For a

---

[7]For instance, one can parametrize a rotation using Euler angles (or a tangent-space representation), which makes the corresponding measurement models nonlinear, but allows treating the variables as unconstrained quantities.

square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, $\text{tr}\,(\boldsymbol{A}) \triangleq \sum_{i=1}^{d} a_{ii}$ denotes the matrix trace. For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times d}$, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle \triangleq \text{tr}\,(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{B}) = \sum_{i=1}^{m} \sum_{j=1}^{d} A_{ij}B_{ij}$ denotes the usual inner product between real matrices. $[\boldsymbol{A}, \boldsymbol{B}]$ and $[\boldsymbol{A}\ \boldsymbol{B}]$ denote the *horizontal* concatenation, while $[\boldsymbol{A}\,;\,\boldsymbol{B}]$ denotes the *vertical* concatenation, for proper $\boldsymbol{A}, \boldsymbol{B}$. For a vector $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|_1$, $\|\boldsymbol{v}\|_2$, and $\|\boldsymbol{v}\|_\infty$ to denote the ℓ_1, ℓ_2, and ℓ_∞ norm of $\boldsymbol{v}$, respectively. For $a \in \mathbb{R}$, the symbol $\lceil a \rceil$ returns the smallest integer $\geq a$. We use $\mathbb{S}^d$ to denote the space of $d \times d$ real symmetric matrices, and $\mathbb{S}^d_+$ (resp. $\mathbb{S}^d_{++}$) to denote the set of matrices in $\mathbb{S}^n$ that are *positive semidefinite* (resp. definite). We also write $\boldsymbol{X} \succeq 0$ (resp. $\boldsymbol{X} \succ 0$) to indicate $\boldsymbol{X}$ is positive semidefinite (resp. definite). We denote with $\mathbb{N}$ the set of natural numbers (nonnegative integers), and for a given $m \in \mathbb{N}$ with $m \geq 1$, we use the notation $[m] \triangleq \{1, \ldots, m\}$ to denote the set of indices from 1 to m. For a finite set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. Finally, $\boldsymbol{1}_\mathcal{A} \in \{0; 1\}^m$ is the indicator vector of the set $\mathcal{A} \subseteq [m]$, whose i-th entry is 1 if $i \in \mathcal{A}$ or zero otherwise.

4.1 Semidefinite Programming

The algorithms discussed in this monograph require solving large semidefinite programs. A semidefinite program is a convex optimization problem and can be solved in polynomial time by off-the-shelf optimization solvers, *e.g.,* [105, 106]. More formally, a *multi-block* semidefinite programming (SDP) problem is an optimization problem in the following primal form [107]:

$$\min_{\boldsymbol{X} \in \mathbb{X}_{\text{sdp}}} \{\langle \boldsymbol{C}, \boldsymbol{X} \rangle \mid \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b},\ \boldsymbol{X} \succeq 0\}, \tag{SDP}$$

where the variable $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_l)$ is a collection of l square matrices (the "blocks") with $\boldsymbol{X}_i \in \mathbb{R}^{d_i \times d_i}$ for $i = 1, \ldots, l$ (conveniently ordered such that $d_1 \geq \ldots \geq d_l$); the domain $\mathbb{X}_{\text{sdp}} \triangleq \mathbb{S}^{d_1} \times \ldots \times \mathbb{S}^{d_l}$ and the constraint $\boldsymbol{X} \succeq 0$ restrict the matrices to be symmetric positive semidefinite. The objective is a linear combination of the matrices in $\boldsymbol{X}$, *i.e.,* $\langle \boldsymbol{C}, \boldsymbol{X} \rangle \triangleq \sum_{i=1}^{l} \langle \boldsymbol{C}_i, \boldsymbol{X}_i \rangle$ (for given matrices $\boldsymbol{C}_i \in \mathbb{S}^{d_i}, i = 1, \ldots, l$). The problem includes independent linear constraints $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$, where:

$$\mathcal{A}(\boldsymbol{X}) \triangleq \left[\sum_{i=1}^{l} \langle \boldsymbol{A}_{i1}, \boldsymbol{X}_i \rangle\,;\, \ldots\,;\, \sum_{i=1}^{l} \langle \boldsymbol{A}_{im}, \boldsymbol{X}_i \rangle\right] \in \mathbb{R}^m, \tag{8}$$

for given matrices $\boldsymbol{A}_{ij} \in \mathbb{S}^{d_i}, i = 1, \ldots, l$, and $j = 1, \ldots, m$, and a given vector $\boldsymbol{b} \in \mathbb{R}^m$.

4.2 Polynomial Optimization and Lasserre's Hierarchy of Moment Relaxations

Our interest towards polynomial optimization stems from the fact that the outlier-robust formulations discussed in Section 1, *i.e.,* (LTS), (MC), and (TLS), can be written as polynomial optimization problems when the measurement model is linear (or, more generally, polynomial) and $\boldsymbol{x}$ belongs to a basic semi-algebraic set. Moreover, as we will see in this section, Lasserre's hierarchy of moment relaxations provides a systematic approach to relax polynomial optimization problems (which in general are hard to solve) into semidefinite programs (which can be solved in polynomial time). Indeed, the tools covered in this section already provide a fairly general template to design outlier-robust estimators for geometric perception and have been used in recent works [30, 33, 1, 34, 35].

Polynomial optimization. Given a vector $\boldsymbol{x} = [x_1\,;\, x_2\,;\, \ldots\,;\, x_{d_x}] \in \mathbb{R}^{d_x}$, a *monomial* in $\boldsymbol{x}$ is a product of x_i's with *nonnegative* integer exponents (for instance $x_1^2 x_5 x_6^3$ is a monomial). The sum of the exponents is called the *degree* of the monomial (*e.g.,* the monomial $x_1^2 x_5 x_6^3$ has degree 6). A real *polynomial* $p(\boldsymbol{x})$ is a finite sum of monomials with real coefficients. We shorthand p in place of $p(\boldsymbol{x})$ when the variable $\boldsymbol{x}$ is clear from the context. The degree of a polynomial p, denoted by $\deg\,(p)$, is the *maximum* degree of its monomials. The ring of real polynomials is denoted by $\mathbb{R}[\boldsymbol{x}]$.

A polynomial optimization problem (POP) is an optimization problem in the form:

$$p^\star \triangleq \min_{\boldsymbol{x}\in\mathbb{R}^{d_x}} \left\{ p(\boldsymbol{x}) \;\middle|\; \begin{array}{l} h_i(\boldsymbol{x}) = 0, i = 1,\dots,l_h \\ g_j(\boldsymbol{x}) \geq 0, j = 1,\dots,l_g \end{array} \right\}, \tag{POP}$$

where $p, h_i, g_j \in \mathbb{R}[\boldsymbol{x}]$. Problem (POP) is hard to solve in general [104] (*e.g.,* hard combinatorial problems with binary constraints $x_i \in \{0,1\}$ can be written in the form (POP) by imposing $x_i^2 = x_i$, $i = 1,\dots,d_x$), but it admits a well-studied convex relaxation that we review below.

Lasserre's hierarchy of moment relaxations [104, 25, 108]. We denote with $[\boldsymbol{x}]_r$ the vector of monomials of degree up to r. For example, if $\boldsymbol{x} = [x_1 \,;\, x_2]$ and $r = 2$, then $[\boldsymbol{x}]_2 = [1 \,;\, x_1 \,;\, x_2 \,;\, x_1^2 \,;\, x_1 x_2 \,;\, x_2^2]$. The dimension of $[\boldsymbol{x}]_r$ is $\underline{d}_r \triangleq \binom{d_x+r}{r}$. With $[\boldsymbol{x}]_r$, we form the so-called *moment matrix* $\boldsymbol{X}_{2r} \triangleq [\boldsymbol{x}]_r[\boldsymbol{x}]_r^\mathsf{T}$. For instance, for $\boldsymbol{x} = [x_1 \,;\, x_2]$ and $r = 2$ (*cf.* $[\boldsymbol{x}]_2$ above)[8]:

$$\boldsymbol{X}_4 \triangleq [\boldsymbol{x}]_2[\boldsymbol{x}]_2^\mathsf{T} = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_1 x_2 & x_2^2 \\ x_1 & x_1^2 & x_1 x_2 & x_1^3 & x_1^2 x_2 & x_1 x_2^2 \\ x_2 & x_1 x_2 & x_2^2 & x_1^2 x_2 & x_1 x_2^2 & x_2^3 \\ x_1^2 & x_1^3 & x_1^2 x_2 & x_1^4 & x_1^3 x_2 & x_1^2 x_2^2 \\ x_1 x_2 & x_1^2 x_2 & x_1 x_2^2 & x_1^3 x_2 & x_1^2 x_2^2 & x_1 x_2^3 \\ x_2^2 & x_1 x_2^2 & x_2^3 & x_1^2 x_2^2 & x_1 x_2^3 & x_2^4 \end{bmatrix}. \tag{9}$$

By construction, $\boldsymbol{X}_{2r}$ is positive semidefinite and has $\operatorname{rank}(\boldsymbol{X}_{2r}) = 1$. Moreover, the set of *unique* entries in $\boldsymbol{X}_{2r}$ is simply $[\boldsymbol{x}]_{2r}$, *i.e.,* the set of monomials of degree up to $2r$ (these monomials appear multiple times in $\boldsymbol{X}_{2r}$, *e.g.,* see $x_1 x_2$ in eq. (9)). Therefore, a key fact is that —*for a suitable matrix $\boldsymbol{A}$*— the linear function $\langle \boldsymbol{A}, \boldsymbol{X}_{2r} \rangle$ *can express any polynomial in $\boldsymbol{x}$ of degree up to $2r$.*[9]

The key idea of Lasserre's hierarchy of moment relaxations is to (i) rewrite (POP) using the moment matrix $\boldsymbol{X}_{2r}$, (ii) relax the (non-convex) rank-1 constraint on $\boldsymbol{X}_{2r}$ (and only enforce $\boldsymbol{X}_{2r} \succeq 0$, which is a convex constraint),[10] and (iii) add redundant constraints that are trivially satisfied in (POP) but still contribute to improving the quality of the relaxation. This leads to a *convex* semidefinite program in the form (SDP) (the interested reader can find a more detailed derivation in Appendix A), which can be conveniently solved in polynomial time:

$$m^\star \triangleq \min_{\boldsymbol{X}\in\mathbb{X}_{\mathsf{sdp}}} \left\{ \langle \boldsymbol{C}, \boldsymbol{X} \rangle \mid \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \; \boldsymbol{X} \succeq 0 \right\}. \tag{LAS$_r$}$$

One can solve the relaxation for different choices of r, leading to a *hierarchy* of convex relaxations; r is typically referred to as the *order* of the relaxation. The importance of Lasserre's hierarchy lies in its stunning theoretical properties: when the set defined by the constraints is explicitly bounded (Assumption 2), then Lasserre's relaxation (LAS$_r$) can be proven to compute a solution to the

---

[8]Contrary to [30], we use the subscript $2r$ instead of r for the moment matrix computed as $[\boldsymbol{x}]_r[\boldsymbol{x}]_r^\mathsf{T}$; we believe this notation is more intuitive (the moment matrix contains monomials up to degree $2r$) and is more consistent with the standard definition of pseudo-distributions, as we will see in Appendix B.

[9]For instance, we can write the polynomial $5x_1^3 + 0.3x_1^4 + 4x_1^2 x_2^2 + x_2^4$ as $\langle \boldsymbol{A}, \boldsymbol{X}_{2r} \rangle$ with:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{5}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{5}{2} & 0 & 0.3 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 \end{bmatrix}. \tag{10}$$

[10]As shown in Appendix B, relaxing the rank constraint converts the moment matrix into a *pseudo-moment matrix*.

original (POP) when $r \to \infty$, and —under further technical conditions— the relaxation computes a solution to (POP) for some finite r [108, 109]. In practice, the dimension of $\boldsymbol{X}_{2r}$ quickly grows for increasing r, hence solving the SDP (LAS_r) is only practical for small relaxation orders r. Therefore, it would be interesting to see if we can compute a good approximation of the solution $\boldsymbol{x}^\star$ of (POP) from the solution of the moment relaxation at some small r. We expand on this point below.

Rounding and a posteriori guarantees. How can we estimate a "good" solution for (POP) and assess its quality using (LAS_r)? Given a solution $\boldsymbol{X}^\star$ to (LAS_r) (again, computable in polynomial time) for some small relaxation order r, one can extract a feasible (but possibly suboptimal) estimate $\hat{\boldsymbol{x}}$ for (POP), using a *rounding* procedure; for instance, we can extract the entries of $\boldsymbol{X}^\star$ corresponding to $\boldsymbol{x}$, which we denote as $\boldsymbol{X}^\star_{[\boldsymbol{x}]}$ (*cf.* (9), where we can extract $\boldsymbol{X}^\star_{[\boldsymbol{x}]} = [x_1 \, ; \, x_2]$ from the first column of the matrix), and then project the corresponding vector to the set $\mathbb{X}$. The rounding procedure is often problem dependent, but it is straightforward to implement in the geometric perception problems considered in this monograph, where projecting to the feasible set (POP) is easy (*e.g.,* given a generic 3×3 matrix, it is easy to project the matrix onto the set of 3D rotations, see [4]). Interestingly, checking the quality of the estimate $\hat{\boldsymbol{x}}$ *a posteriori* (*i.e.,* after solving the moment relaxation) is also easy: if we call $\hat{p} \triangleq p(\hat{\boldsymbol{x}})$ the objective attained by $\hat{\boldsymbol{x}}$ in (POP), it holds:

$$m^\star \leq p^\star \leq \hat{p}, \tag{11}$$

where the first inequality follows from the fact that (LAS_r) is a relaxation, while the second follows from the fact that $p^\star$ is the optimal (lowest) cost over the feasible set of (POP) and $\hat{\boldsymbol{x}}$ is feasible for (POP). Since we can compute $m^\star$ after solving the relaxation (LAS_r) and we can also compute $\hat{p}$ after rounding, we can use (11) to bound the suboptimality of $\hat{\boldsymbol{x}}$ (*i.e.,* from (11), we can bound the suboptimality gap of $\hat{\boldsymbol{x}}$ as $\hat{p} - p^\star \leq \hat{p} - m^\star$) and understand how far is $\hat{\boldsymbol{x}}$ from an optimal solution. Moreover, if $\hat{p} = m^\star$, the inequalities (11) become tight, and we can conclude that $\hat{p} = p^\star$ and $\hat{\boldsymbol{x}}$ is indeed an optimal solution. This a posteriori checks are at the basis of the certifiable algorithms proposed in robotics and vision [30, 33, 1, 34, 35], which first reformulate robust estimation as a (POP) and then apply the following algorithmic workflow:

$$(\text{POP}) \xrightarrow{\text{moment relaxation}} (\text{LAS}_r) \xrightarrow{\text{SDP solver}} (\boldsymbol{X}^\star, m^\star) \xrightarrow{\text{rounding}} (\hat{\boldsymbol{x}}, \hat{p}) \xrightarrow{\text{certification}} \hat{p} \stackrel{?}{=} m^\star, \tag{12}$$

where the last step is to certify optimality of $\hat{\boldsymbol{x}}$ whenever $\hat{p} = m^\star$. In practice, what is making the works [30, 33, 1, 34, 35] (as well as the previous work on outlier-free estimation, *e.g.,* [7, 110, 13, 111, 46]) compelling is the empirical observation that the moment relaxation is *tight* (meaning $\hat{p} = m^\star$) in many practical problems for $r = 1$ or $r = 2$, hence the workflow above allows computing optimal solutions to (POP) efficiently in practice. Moreover, the certification step can be directly used in practice, *e.g.,* a robot can trust an estimate if it is certified as optimal or discard it (or at least handle it more carefully) if no optimality certificate is obtained.

What about a priori guarantees? The a posteriori guarantees above can only be obtained for a given problem and input data after solving the corresponding relaxation, and for a specific choice of rounding. Moreover, the empirical observation that the relaxation is often tight seems to be a *deus ex machina* and unexpectedly solve a provably hard problem. Therefore in this monograph we are interested in *a priori guarantees*. How can we theoretically justify the empirical performance of the moment relaxation for geometric perception? Can we characterize the input data for which we expect to obtain good relaxations and rounded estimates close to the ground-truth variable we

15

are trying to estimate? These questions are of practical relevance, since answering these questions (i) would enable a better understanding of the conditions under which the perception system of a robot is expected to work well, and (ii) would allow the design of novel perception front-ends that can produce measurements that are more likely to lead to good estimates.

Providing a priori guarantees would be easy for very large relaxation orders, *i.e.,* $r \to \infty$, since in this case the solution of (LAS_r) can be proven to retrieve the solution of (POP) (see Appendix A). However, here we want to provide a priori guarantees for very small r. In robotics and vision, such guarantees have only appeared for specific problems [7, 46, 34, 36]. However, it would be desirable to have a more general language to discuss properties of moment relaxations for small relaxation orders r. Luckily, the sum-of-squares proofs system, described below, provides such a language.

4.3 Sum-of-Squares Proofs

Sum-of-squares (sos) proofs provide an advanced way to reason about polynomial constraints and to infer properties of the moment relaxation introduced above. Here we want to give some intuition about the sos proof system, and we postpone a more formal introduction to Appendix C.

Why do we need to reason about polynomial constraints? Let us start with some intuition and motivation, before formalizing the concept of sos proof. Assume that we rephrased one of the problems presented in the introduction, *e.g.,* (TLS), as a (POP) and obtained the corresponding moment relaxation (LAS_r). We then solved (LAS_r) to obtain a solution matrix $\boldsymbol{X}^\star$ (later, we are going to call this object a "pseudo-moment matrix"). Now, we would like to infer that $\boldsymbol{X}^\star$ satisfies some property of interest; for instance, in our estimation problems we may want to ensure that for some suitable linear function $\mathcal{L}(\cdot)$, the following holds:

$$\|\mathcal{L}(\boldsymbol{X}^\star) - \boldsymbol{x}^\circ\|_2^2 \leq \eta^2, \tag{13}$$

which states that the estimate $\mathcal{L}(\boldsymbol{X}^\star)$ computed from $\boldsymbol{X}^\star$ is within a distance η from the ground-truth $\boldsymbol{x}^\circ$. The sos proofs system does exactly that: it provides a systematic way to conclude that the solution of the relaxation of a system of polynomial constraints, such as the constraints in (TLS) (let us call these polynomial constraints $\mathcal{A}$) also satisfies a desired polynomial relation, *e.g.,* (13) (let us call this polynomial constraint $\mathcal{B}$). The key idea is that if we can provide an *sos proof* that $\mathcal{A}$ "implies" $\mathcal{B}$, a novel type of proof that we introduce below, then a moment relaxation of $\mathcal{A}$ will also satisfy $\mathcal{B}$. Note that (13) is only an example of implication we might be attempting to prove, while in general, we might try to prove other (polynomial-expressible) properties of the moment relaxation.

What is an sos proof? First of all, we recall that a polynomial $p(\boldsymbol{x})$ is *sum-of-squares* (sos) if there exist polynomials $q_1(\boldsymbol{x}), \ldots, q_t(\boldsymbol{x})$ such that $p = q_1^2 + \ldots + q_t^2$. Now consider a system of polynomial constraints $\mathcal{A}(\boldsymbol{x}) = \{f_1(\boldsymbol{x}) \geq 0, f_2(\boldsymbol{x}) \geq 0, \ldots, f_m(\boldsymbol{x}) \geq 0\}$ for some given polynomials $f_i(\boldsymbol{x})$, $i \in [m]$, and the inequality $g(\boldsymbol{x}) \geq 0$ (for some polynomial $g(\boldsymbol{x})$). We are then interested in defining a proof that $\mathcal{A}(\boldsymbol{x})$ implies $g(\boldsymbol{x}) \geq 0$, *i.e.,* any $\boldsymbol{x}$ that satisfies $\mathcal{A}(\boldsymbol{x})$ is such that $g(\boldsymbol{x}) \geq 0$.

Definition 3 (Sum-of-squares proof)**.** *Given a system of polynomial constraint $\mathcal{A}$ and a polynomial g, a sum-of-square (sos) proof that the system $\mathcal{A}$ implies $g \geq 0$ consists of sum-of-squares polynomials $\{p_\mathcal{S}\}_{\mathcal{S} \subseteq [m]}$ such that:*

$$g = \sum_{\mathcal{S} \subseteq [m]} p_\mathcal{S} \cdot \prod_{i \in \mathcal{S}} f_i. \tag{14}$$

We say that the proof has degree k if for every $\mathcal{S} \subseteq [m]$, $deg\left(p_\mathcal{S} \cdot \prod_{i \in \mathcal{S}} f_i\right) \leq k$ where $deg(\cdot)$ denotes

the degree of a polynomial. We use the notation:

$$\mathcal{A}(\boldsymbol{x}) \left|\frac{k}{\boldsymbol{x}} \{g(\boldsymbol{x}) \geq 0\} \qquad or \qquad \{f_i(\boldsymbol{x}) \geq 0, \ldots, f_m(\boldsymbol{x}) \geq 0\} \right|\frac{k}{\boldsymbol{x}} \{g(\boldsymbol{x}) \geq 0\} \tag{15}$$

to denote that there is a proof of degree at most k of the fact that $\mathcal{A} = \{f_i(\boldsymbol{x}) \geq 0, \ldots, f_m(\boldsymbol{x}) \geq 0\}$ implies $g \geq 0$ (i.e., any $\boldsymbol{x}$ that satisfies $\mathcal{A}(\boldsymbol{x})$ is such that $g(\boldsymbol{x}) \geq 0$). We omit the variables and write $\mathcal{A}(\boldsymbol{x}) \left|\frac{k}{} \{g(\boldsymbol{x}) \geq 0\}\right.$, when they are clear from the context. Moreover, we write

$$\left|\frac{k}{\boldsymbol{x}} \{g(\boldsymbol{x}) \geq 0\}\right. \tag{16}$$

if there is a sum-of-squares proof that $g(\boldsymbol{x}) \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^{d_x}$ (i.e., $g(\boldsymbol{x})$ is sum-of-squares).

From eq. (14), it is clear why the polynomials $p_{\mathcal{S}}$ are a "proof" of $g \geq 0$ for any $\boldsymbol{x}$ satisfying $\mathcal{A}$: for any $\boldsymbol{x} \in \mathcal{A}$, $\prod_{i \in \mathcal{S}} f_i \geq 0$ by definition, hence if we can write g as the product of a sum-of-squares (hence non-negative) polynomial and $\prod_{i \in \mathcal{S}} f_i$, we automatically prove that $g \geq 0$ whenever $\boldsymbol{x} \in \mathcal{A}$. We note that the existence of an sos proof is a *sufficient* condition for the fact that $g(\boldsymbol{x}) \geq 0$ whenever $\mathcal{A}(\boldsymbol{x})$ is satisfied. However, it is not a necessary condition, in that there might be valid relations that cannot be proved using sos proofs. For instance, while the polynomial $p(\boldsymbol{x}) = x_1^4 x_2^2 + x_1^2 x_2^4 + 1 - 3x_1^2 x_2^2$ (the Motzkin polynomial, see [48, p. 59]) is such that $p(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^2$, it is not possible to develop an sos proof for such a fact, since $p(\boldsymbol{x})$ is not sos [48, p. 59]. In other words, the sos proof system is more stringent than the traditional proofs we might be used to, and properties that hold with traditional proofs might not hold in the sos sense. At the same time, if we are able to derive an sos proof we can obtain strong guarantees for our moment relaxations, and existing results reassure us that all relevant properties of moment relaxations can be proven via sos proofs (see Appendix C).

How to derive an sos proof? Sum-of-squares provide a proof system to reasons about polynomial constraints. For instance, imagine that our goal is to derive a proof that $\mathcal{A}$ implies $g \geq 0$. In a traditional mathematical proof system, we might first prove that $\mathcal{A}$ implies $g' \geq 0$ (for some other polynomial g') and that $g \geq g'$, to finally conclude that $\mathcal{A}$ implies $g \geq 0$. Similarly, the sos proof system provides a systematic mechanism to derive this chain of implications, but *with more stringent rules* compared to the ones we are typically used to in robotics and vision. For instance, we have already observed the fact that $p(\boldsymbol{x}) \geq 0$ for some degree k polynomial does not necessarily imply that there is a sum-of-squares proof $\left|\frac{k}{\boldsymbol{x}} \{p(\boldsymbol{x}) \geq 0\}\right.$. Similarly, for some polynomial constraints $\mathcal{A}$ and polynomials $g'(\boldsymbol{x})$ and $g(\boldsymbol{x})$ such that $g(\boldsymbol{x}) \geq g'(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^{d_x}$, the fact that $\mathcal{A} \left|\frac{k}{\boldsymbol{x}} \{g'(\boldsymbol{x}) \geq 0\}\right.$ does not necessarily imply that $\mathcal{A} \left|\frac{k}{\boldsymbol{x}} \{g(\boldsymbol{x}) \geq 0\}\right.$, since the latter fact might not admit a sum-of-squares proof. In this sense, the sos proof system is more restrictive than the typical algebraic manipulation we are used to. Fortunately, previous work provides a toolkit of inference rules that can be used to correctly reason in the sos proof system. We collect a set of "sos rules" in Appendix C, mostly drawing from [37, 15, 112, 113, 114].

How to use sos proofs? Assume we were able to derive a proof that $\mathcal{A}(\boldsymbol{x}) \left|\frac{k}{\boldsymbol{x}} \{g(\boldsymbol{x}) \geq 0\}\right.$. Then, the key fact that we mention here informally, and formalize in Appendix C, is that any pseudo-moment matrix that satisfies a moment relaxation of $\mathcal{A}$ also satisfies a moment relaxation of $g \geq 0$.[11] This fact will be instrumental in proving that the solution of the moment relaxations of (LTS), (TLS), and (MC) have to satisfy some desirable properties, and will be key to deriving the error bounds we present below. The sos proof system has found extensive applications in algorithmic

---

[11]We will need some extra notation to make this claim more precise, and postpone those details to Appendix C.

statistics (tracing back to the seminal paper [115]), leading to the so called "proof to algorithms" paradigm, where the proof system directly suggests tractable algorithms to solve a problem.

The reader should be able to follow the rest of this monograph without reading the appendices, whose material is mostly useful to support the technical proofs and for a more rigorous introduction to the material informally reviewed in this section. At the same time, we invite the reader to use Appendix A to Appendix C as an accessible introduction to the world of moment relaxations and sos proofs: in those appendices, we attempt to bridge the advanced presentation that is typically found in statistics papers and books with the more familiar optimization lens used in robotics and vision.

5   Estimation Contracts: Problem Statement

This monograph is concerned with the following problems.

Problem 1 (Outlier-robust estimation in geometric perception). *Estimate $\boldsymbol{x}^\circ \in \mathbb{X}$ (where $\mathbb{X}$ is a basic semi-algebraic set) given n measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, such that a subset of $\alpha\, n$ measurements (the* inliers*) follows the measurement model:*

$$\boldsymbol{y}_i = \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ + \boldsymbol{\epsilon}, \quad with \quad \boldsymbol{y}_i \in \mathbb{R}^{d_y} \quad and \quad \boldsymbol{x}^\circ \in \mathbb{X} \subseteq \mathbb{R}^{d_x}, \tag{17}$$

with $\|\boldsymbol{\epsilon}\|_2 \le \bar{c}$ (for a given noise bound $\bar{c}$) and the remaining $\beta\, n$ measurements (the outliers*, with $\beta = 1 - \alpha$) are arbitrary (and possibly adversarially chosen).*

Problem 2 (Estimation contracts). *For each algorithm developed to solve Problem 1, provide conditions on the inliers such that the resulting estimate is guaranteed to be close to the ground truth $\boldsymbol{x}^\circ$.*

As discussed in Section 3 (and further stressed in Section 9 below), Problem 1 arises in many geometric perception applications. Our main focus in this monograph will be on designing estimation contracts (Problem 2): rather than proposing radically new algorithms for Problem 1, we review existing algorithms (possibly with small modifications), and then derive suitable conditions under which those algorithms are guaranteed to return good estimates.

Section 6 below studies the low-outlier case, where $\beta \ll 0.5$. There, we review algorithms to attack Problem 1 based on moment relaxations of (LTS), (MC), and (TLS). Then, for each algorithm, we provide estimation contracts.

Section 7 studies the high-outlier case, where $\beta \gg 0.5$. In such a case, a sound algorithm must return multiple estimates and our estimation contracts derive conditions under which at least one of the returned estimates is close to the ground truth. The results we present are an adaptation to the geometric perception setup of the recent work on *list-decodable regression* by Karmalkar *et al.* [37].

6   Estimation Contracts for Low Outlier Rates

This section provides estimation contracts for (LTS), (MC), and (TLS) for problems with low-outlier rates (*i.e.,* $\beta \ll 0.5$). We start by deriving some naive contracts for (MC) and (TLS) in Section 6.1: these naive results only use basic manipulations and do not rely on any of the machinery presented above; at the same time, their derivation shares several insights with the more advanced contracts we provide in the subsequent sections and motivates the need for the machinery in Section 4. Then, Section 6.2.1, Section 6.2.2, and Section 6.2.3 present more advanced estimation contracts based on sos proofs for (LTS), (MC), and (TLS), respectively.

6.1 A Posteriori (Naive) Estimation Contracts for (MC) and (TLS)

Here we develop two simple results for (MC) and (TLS) that quantify the distance of an optimal solution of the two problems from the ground-truth parameter $\boldsymbol{x}^\circ$. These results are straightforward to prove, but have several shortcomings that we discuss in Remark 7 at the end of the section.

Before presenting the results in this section, we need to remark that in all the estimation problems considered in Section 3, we cannot reconstruct the unknown $\boldsymbol{x}^\circ$ from a single measurement, but we rather need a sufficiently large subset of measurements.

Definition 4 (Nondegenerate and minimal measurement set). *A set $\mathcal{J}$ of measurements is* nondegenerate *if the following optimization problem admits a unique solution:*

$$\min_{\boldsymbol{x} \in \mathbb{X}} \sum_{i \in \mathcal{J}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x} \right\|_2^2. \tag{18}$$

A nondegenerate set of minimal size $\bar{d}$ is called a minimal *measurement set.*

Characterizations of the minimal sets for common geometric perception problems are well known in the literature. Indeed, a subfield of computer vision is specifically concerned with the design of *minimal solvers*, which compute an estimate $\boldsymbol{x}$ from a minimal set of measurements.[12] For instance, the rotation search problem and the 3D registration problem in Section 3 require at least $\bar{d} = 3$ non-collinear 3D point measurements for the resulting estimate in (18) to be unique.

The following proposition provides our first estimation contract, which establishes when an optimal solution of (MC) is close to the ground truth $\boldsymbol{x}^\circ$.

Proposition 5 (Low-outlier case: a posteriori estimation contract for (MC)). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and assume the measurement set contains $\alpha\, n \geq \frac{n+\bar{d}}{2}$ inliers, where $\bar{d}$ is the size of a minimal set. Moreover, assume that every subset of $\bar{d}$ inliers is nondegenerate. Then, for any integer $d_\mathcal{J}$ such that $\bar{d} \leq d_\mathcal{J} \leq (2\alpha - 1)n$, an optimal solution $\boldsymbol{x}_{\mathsf{MC}}$ of (MC) satisfies*

$$\| \boldsymbol{x}_{\mathsf{MC}} - \boldsymbol{x}^\circ \|_2 \leq \frac{2\sqrt{d_\mathcal{J}}\, \bar{c}}{\min_{\mathcal{J} \subset \mathcal{I}_{\mathsf{MC}}, |\mathcal{J}| = d_\mathcal{J}} \sigma_{\min}(\boldsymbol{A}_\mathcal{J})}, \tag{19}$$

where $\mathcal{I}_{\mathsf{MC}}$ is the set of inliers selected by (MC), $\boldsymbol{A}_\mathcal{J}$ is the matrix obtained by horizontally stacking all submatrices $\boldsymbol{A}_i$ for all $i \in \mathcal{J}$, and $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix. Moreover, if the inliers are noiseless, i.e., $\boldsymbol{\epsilon} = \boldsymbol{0}$ in eq. (17), and $\bar{c} = 0$, then $\boldsymbol{x}_{\mathsf{MC}} = \boldsymbol{x}^\circ$.

We report the proof here (while all other technical proofs are postponed to the appendix) since the structure of the proof is quite enlightening in its simplicity. Indeed, we will see that this simple proof shares many insights with the proofs of more advanced results presented later in this monograph.

Proof. Given an optimal solution $(\boldsymbol{x}_{\mathsf{MC}}, \boldsymbol{\omega}_{\mathsf{MC}})$ of problem (MC), let us call $\mathcal{I}$ and $\mathcal{I}_{\mathsf{MC}}$ the true inliers and the set of measurements selected by (MC) (*i.e.*, $\mathcal{I}_{\mathsf{MC}} \triangleq \{i \in [n] : \omega_{\mathsf{MC},i} = 1\}$), respectively. Recall that the proposition assumes $|\mathcal{I}| = \alpha\, n \geq \frac{n+\bar{d}}{2}$. We prove the result in two steps.

(i) The solution of (MC) captures enough inliers. We denote with $\boldsymbol{1}_\mathcal{I}$ the indicator vector of the set $\mathcal{I}$, *i.e.*, the n-vector that has the i-th entry equal to 1 if $i \in \mathcal{I}$ or zero otherwise. We then

---

[12]The popularity of minimal solvers stems from their extensive use in outlier rejection schemes, such as RANSAC [51].

note that $(\boldsymbol{x}^\circ, \mathbf{1}_{\mathcal{I}})$ is a feasible solution for (MC) and attains a cost $\alpha\,n$. Therefore, by optimality of $(\boldsymbol{x}_{\mathsf{MC}}, \boldsymbol{\omega}_{\mathsf{MC}})$, it follows $|\mathcal{I}_{\mathsf{MC}}| \geq \alpha\,n$. Since $|\mathcal{I}| = \alpha\,n$ and $|\mathcal{I}_{\mathsf{MC}}| \geq \alpha\,n$ then:

$$|\mathcal{I} \cap \mathcal{I}_{\mathsf{MC}}| \overset{\overbrace{\text{sets overlap in } [n]}}{\geq} (2\alpha - 1)n \overset{\overbrace{\text{using } \alpha\,n \geq \frac{n+\bar{d}}{2}}}{\geq} \bar{d}, \tag{20}$$

i.e., the measurements selected by (MC) must include at least $(2\alpha - 1)n \geq \bar{d}$ inliers.

   (ii) The inliers captured by (MC) bound the estimation error. Let us pick any subset $\mathcal{J}$ of the set $\mathcal{I} \cap \mathcal{I}_{\mathsf{MC}}$, such that $|\mathcal{J}| = d_{\mathcal{J}}$, for any choice of integer $d_{\mathcal{J}}$ such that $\bar{d} \leq d_{\mathcal{J}} \leq (2\alpha - 1)n$. Note that both $\boldsymbol{x}^\circ$ and $\boldsymbol{x}_{\mathsf{MC}}$ are feasible for (MC) over $\mathcal{J}$, hence:

$$\max_{i \in \mathcal{J}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}^\circ \right\|_2^2 \leq \bar{c}^2 \implies \sum_{i \in \mathcal{J}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}^\circ \right\|_2^2 \leq d_{\mathcal{J}}\,\bar{c}^2 \implies \left\| \boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}^\circ \right\|_2 \leq \sqrt{d_{\mathcal{J}}}\,\bar{c} \tag{21}$$

$$\max_{i \in \mathcal{J}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}_{\mathsf{MC}} \right\|_2^2 \leq \bar{c}^2 \implies \sum_{i \in \mathcal{J}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}_{\mathsf{MC}} \right\|_2^2 \leq d_{\mathcal{J}}\,\bar{c}^2 \implies \left\| \boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}_{\mathsf{MC}} \right\|_2 \leq \sqrt{d_{\mathcal{J}}}\,\bar{c}, \tag{22}$$

where $\boldsymbol{y}_{\mathcal{J}}$ and $\boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}}$ vertically stack all the measurements $\boldsymbol{y}_i$ and matrices $\boldsymbol{A}_i^{\mathsf{T}}$ for $i \in \mathcal{J}$.

   Now we observe that the mismatch between $\boldsymbol{x}_{\mathsf{MC}}$ and $\boldsymbol{x}^\circ$ after multiplying by $\boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}}$ must be small:

$$\left\| \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} (\boldsymbol{x}_{\mathsf{MC}} - \boldsymbol{x}^\circ) \right\|_2 \overset{\overbrace{\text{adding and subtracting } \boldsymbol{y}_{\mathcal{J}}}}{=} \left\| (\boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}^\circ) - (\boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}_{\mathsf{MC}}) \right\|_2 \tag{23}$$

$$\overset{\overbrace{\text{triangle inequality}}}{\leq} \left\| \boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}^\circ \right\|_2 + \left\| \boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}_{\mathsf{MC}} \right\|_2 \overset{\overbrace{\text{using (21)-(22)}}}{\leq} 2\sqrt{d_{\mathcal{J}}}\,\bar{c}. \tag{24}$$

Recalling that for a matrix $\boldsymbol{M}$ and vector $\boldsymbol{v}$, $\|\boldsymbol{M}\boldsymbol{v}\|_2 \geq \sigma_{\min}(\boldsymbol{M})\,\|\boldsymbol{v}\|_2$, where $\sigma_{\min}(\boldsymbol{M})$ is the smallest singular value of $\boldsymbol{M}$:

$$\left\| \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} (\boldsymbol{x}_{\mathsf{MC}} - \boldsymbol{x}^\circ) \right\|_2 \geq \sigma_{\min}(\boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}})\,\|\boldsymbol{x}_{\mathsf{MC}} - \boldsymbol{x}^\circ\|_2. \tag{25}$$

Combining (25) and (24):

$$\|\boldsymbol{x}_{\mathsf{MC}} - \boldsymbol{x}^\circ\|_2 \leq \frac{2\sqrt{d_{\mathcal{J}}}\,\bar{c}}{\sigma_{\min}(\boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}})}. \tag{26}$$

Since we do not know which subset $\mathcal{J}$ was selected by (MC), we choose the set $\mathcal{J}$ of cardinality $d_{\mathcal{J}}$ attaining the smallest singular value across the set of inliers selected by (MC), yielding the desired result for the case of noisy inliers. In the case of noiseless inliers and $\bar{c} = 0$, (21)-(22) hold exactly, *i.e.,* $\boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}^\circ = \mathbf{0}$ and $\boldsymbol{y}_{\mathcal{J}} - \boldsymbol{A}_{\mathcal{J}}^{\mathsf{T}} \boldsymbol{x}_{\mathsf{MC}} = \mathbf{0}$ and both $\boldsymbol{x}^\circ$ and $\boldsymbol{x}_{\mathsf{MC}}$ attain the minimum (with zero cost in this case) of (18). However, from the non-degeneracy assumption, problem (18) admits a unique minimizer, hence $\boldsymbol{x}_{\mathsf{MC}} = \boldsymbol{x}^\circ$, which concludes the proof. ∎

   The proof of Proposition 5 involves two steps: (i) we proved that the solution of (MC) must capture a sufficient number of true inliers (*i.e.,* there must be enough overlap between the set of measurements selected by (MC) and the true inliers), (ii) the fact that the solution $\boldsymbol{x}_{\mathsf{MC}}$ has to be consistent with the true inliers forces the estimation error to be small. In the noiseless case ($\bar{c} = 0$),
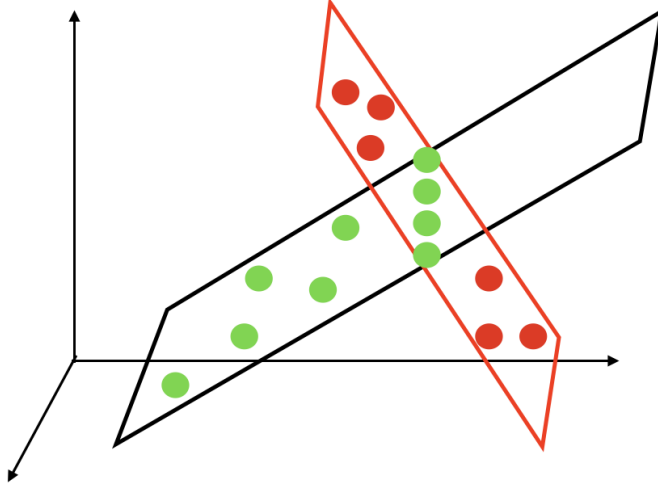
Figure 1: Plane fitting with 9 inliers (green 3D points) and 6 outliers (red 3D points). True plane is shown in black, while the plane computed by (MC) is shown in red. When the estimation contract in Proposition 5 is violated, (MC) may fail to recover the variable $\boldsymbol{x}^\circ$ even when given $\frac{n+\bar{d}}{2}$ inliers.

the proposition predicts that $\boldsymbol{x}_{\mathsf{MC}} = \boldsymbol{x}^\circ$ as long as there are at least $\frac{n+\bar{d}}{2}$ (nondegenerate) inliers.[13] The noisy bound in (19) holds for any choice of integer $d_{\mathcal{J}}$ between $\bar{d}$ and $(2\alpha - 1)n$. As we will see in Section 8, larger values of $d_{\mathcal{J}}$ lead to better bounds but are harder to compute (due to the need to search over all possible subsets of measurements of size $d_{\mathcal{J}}$).

Fig. 1 stresses the key role of nondegeneracy in the estimation contract in Proposition 5. The figure shows a simple linear regression problem in 3D where we need to fit a plane given 3D points belonging to the plane. In particular, we are given a set of 15 measurements with 9 inliers and 6 outliers. While the set of inliers has size $\frac{n+\bar{d}}{2} = 9$ (which satisfies one requirement in the proposition), there are subsets of degenerate points (in particular, the 4 collinear points at the intersection between the two planes). In this case, (MC) will produce the estimate in red, which is far from the ground truth (black plane). Intuitively, degenerate sets of measurements can be easily "stolen" by an adversary since they are compatible with multiple estimates of $\boldsymbol{x}$ (including estimates far from $\boldsymbol{x}^\circ$).

A result similar to Proposition 5 can be proven for (TLS).

---

[13]While still easy to prove, in the case of noiseless inliers and $\bar{c} = 0$, the result $\boldsymbol{x}_{\mathsf{MC}} = \boldsymbol{x}^\circ$ does not directly follow as a consequence of the noisy bound $\|\boldsymbol{x}_{\mathsf{MC}} - \boldsymbol{x}^\circ\|_2 \leq \frac{2\sqrt{d_{\mathcal{J}}}\,\bar{c}}{\min_{\mathcal{J}\subset\mathcal{I}_{\mathsf{MC}},|\mathcal{J}|=d_{\mathcal{J}}}\sigma_{\min(\boldsymbol{A}_{\mathcal{J}})}}$: in the noiseless case the singular value $\sigma_{\min(\boldsymbol{A}_{\mathcal{J}})}$ might become zero for any subset $\mathcal{J}$ (since the ground truth might be in the null space of $\boldsymbol{A}_i^\mathsf{T}$, cf. (6) and (7)), therefore both the numerator (in particular, $\bar{c}$) and the denominator of the noisy upper bound go to zero.

Proposition 6 (Low-outlier case: a posteriori estimation contract for (TLS)). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and denote with γ° the squared residual error of the ground truth $\boldsymbol{x}^\circ$ over the set of inliers $\mathcal{I}$, i.e., $\gamma^\circ \triangleq \sum_{i \in \mathcal{I}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ \right\|_2^2$. Moreover, assume the measurement set contains at least $\frac{n+\bar{d}}{2} + \frac{\gamma^\circ}{\bar{c}^2}$ inliers, where $\bar{d}$ is the size of a minimal set, and that every subset of $\bar{d}$ inliers is nondegenerate. Then, for any integer $d_{\mathcal{J}}$ such that $\bar{d} \leq d_{\mathcal{J}} \leq (2\alpha - 1)n - \frac{\gamma^\circ}{\bar{c}^2}$, an optimal solution $\boldsymbol{x}_{\mathsf{TLS}}$ of (TLS) satisfies*

$$\|\boldsymbol{x}_{\mathsf{TLS}} - \boldsymbol{x}^\circ\|_2 \leq \frac{2\sqrt{d_{\mathcal{J}}}\ \bar{c}}{\min_{\mathcal{J} \subset \mathcal{I}_{\mathsf{TLS}}, |\mathcal{J}| = d_{\mathcal{J}}} \sigma_{\min}(\boldsymbol{A}_{\mathcal{J}})}, \tag{27}$$

where $\mathcal{I}_{\mathsf{TLS}}$ is the set of inliers selected by (TLS), $\boldsymbol{A}_{\mathcal{J}}$ is the matrix obtained by horizontally stacking all submatrices $\boldsymbol{A}_i$ for all $i \in \mathcal{J}$, and $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix. Moreover, if the inliers are noiseless, i.e., $\boldsymbol{\epsilon} = \boldsymbol{0}$ in eq. (17), and for a sufficiently small $\bar{c} > 0$, $\boldsymbol{x}_{\mathsf{TLS}} = \boldsymbol{x}^\circ$.

While Proposition 5 and Proposition 6 allow us to gain intuition about the problem, the applicability of these estimation contracts is limited, as we discuss below.

Remark 7 (A posteriori estimation contracts). *The estimation contracts in Proposition 5 and Proposition 6 refer to the optimal solutions of (MC) and (TLS), respectively. However, both problems are NP-hard [28, 29], hence computing optimal solutions is difficult in general. We can still apply these contracts as follows. Assume that we solve (MC) and (TLS) using the moment relaxation described in Section 4 (see also [30]) and that the relaxation produces a certifiably optimal result (cf. the workflow (12)); then, such a certifiably optimal solution would still enjoy the guarantees in Proposition 5 and Proposition 6. However, since we can only check optimality a posteriori (and only apply these contracts when the relaxation is tight), we call these contracts "a posteriori" estimation contracts.*

In practice, we would like to have more general performance guarantees regardless of the tightness of the moment relaxation. Luckily, the sos proof machinery we introduced in Section 4 does exactly that: it allows inferring properties of the solution of moment relaxations, regardless of its tightness. Using such a machinery, we can derive "a priori" estimation contracts as shown below.

Remark 8 (Approximation Bounds for Inapproximable Problems?). *We have already observed that common robust estimators have been proven to be NP-hard [20, 29], and recent results [28, 29] have concluded that (MC) and (TLS) are inapproximable, i.e., it is not possible to find constant-factor approximation bounds in the worst case. Therefore, our pursuit of a priori performance guarantees might seem to start off on the wrong foot. The key insight behind the results we are going to discuss below is that while (MC) and (TLS) are intractable in the worst case (i.e., there are unlucky problem instances where polynomial-time algorithms cannot retrieve "good" solutions), we can identify properties of the input data (e.g., hypercontractivity, anti-concentration, see Section 6.2 below) that rule out those worst-case instances and enable the computation of performance bounds.*

6.2 A Priori Estimation Contracts for (LTS), (MC), and (TLS)

In the rest of this monograph we develop "a priori" estimation contracts, which establish conditions on the input data such that certain algorithms based on moment relaxations of (LTS), (MC), and (TLS) are able to compute an estimate for Problem 1 which is provably close to the ground truth $\boldsymbol{x}^\circ$.

The philosophy is quite different from the typical work done in robotics and vision. While in those fields, one is typically concerned with the distribution of the measurement noise $\boldsymbol{\epsilon}$ in (3), here

we are concerned with both the measurement noise and the distribution producing the matrices $\boldsymbol{A}_i$ in (3).[14] One for instance might assume the entries of the measurement matrices to be sampled from a zero-mean Gaussian and study the behavior of an outlier-robust estimator. The importance of these matrices for outlier-robust estimation should already be apparent from the statements of Proposition 5 and Proposition 6. In the following, we write $\boldsymbol{A}$ to denote a matrix random variable, while $\boldsymbol{A}_i$, $i \in [n]$, are the realizations of such a random variable. Moreover, we denote with $\widetilde{\mathcal{I}}$ the distribution producing the matrices $\boldsymbol{A}_i$, $i \in [n]$ for the inlier measurements (note the tilde, which we use to distinguish the distribution of the inliers from the inlier set $\mathcal{I} \subseteq [n]$, *cf.* Proposition 5).

In this context, we hope to make the least restrictive assumptions on $\widetilde{\mathcal{I}}$; in other words, we would like our estimators to be guaranteed to work well for a broad class of distributions that generate the input data. Moreover, we want to restrict any assumptions to the inliers, while the mechanism generating outliers can be arbitrary. In particular we will consider two large families of distributions: certifiable hypercontractive and certifiable anti-concentrated distributions, which we introduce below. These definitions are based on [15, 37], but we extend both to be defined over matrices.

Definition 9 (Certifiable hypercontractivity for matrices, adapted from [15]). *For a function $C : [k] \mapsto \mathbb{R}_+$, we say that a distribution $\widetilde{\mathcal{I}}$ over matrices $\boldsymbol{A}$ is k-certifiably C-hypercontractive if for every $t \leq k/2$, there is a degree k sum-of-squares proof of the following inequality in variable $\boldsymbol{v}$*

$$\mathbb{E}_{\widetilde{\mathcal{I}}}\left[\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^{2t}\right] \leq C(t)^t \left(\mathbb{E}_{\widetilde{\mathcal{I}}}\left[\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2\right]\right)^t. \tag{28}$$

We say that a set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably C-hypercontractive if the uniform distribution over the set is k-certifiably C-hypercontractive, i.e.,

$$\left(\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}}\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2^{2t}\right) \leq C(t)^t \left(\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}}\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2^2\right)^t. \tag{29}$$

We remark that certifiable hypercontractivity not only requires (28) (or (29)) to be satisfied, but also requires this fact to have a sum-of-squares proof (see Section 4). Hypercontractivity essentially requires controlling the moments of $\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2$, or, in the case of (29), it bounds the $2t$-norm of the vector with entries $\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2$ with respect to the 2-norm. In the proofs, this property will be used to bound terms such as $\left\|\boldsymbol{A}^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}^\circ)\right\|_2^2$ for some $\boldsymbol{x}$, an aspect that was also key to proving Proposition 5 and Proposition 6 (*cf.* with eq. (25)). What makes certifiable hypercontractivity interesting is the fact that a large class of distributions satisfies this property, including Gaussians, product of distributions with sub-gaussian marginals, and uniform distributions over the boolean hypercube [116]. Moreover, in Section 8 we show that this property is empirically satisfied in many real problems. Finally, we note that, given a set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, we can easily check if (29) is satisfied (for a given t), since one can check if the polynomial expression (29) (in variable $\boldsymbol{v}$) is sum-of-squares in polynomial time using semidefinite programming (see Section 8, Appendix C, and [48, chapter 3]).

While we use hypercontractivity in Theorem 12, in most of the monograph, we will use a stronger assumption on the inlier distribution, known as *certifiable anti-concentration*. Since this property is more involved, we first introduce the concept of anti-concentration, and then provide its sos counterpart, *i.e.,* certifiable anti-concentration. Both Definition 10 and Definition 11 are based on [37] and extend the corresponding definitions to work with matrices (rather than vectors).

---

[14]While these matrices are considered given in robotics and computer vision applications, we can imagine they are generated by sampling from a distribution.

Definition 10 (Anti-concentration for matrices, adapted from [37]). *A zero-mean random matrix* $\boldsymbol{A} \in \mathbb{R}^{d_x \times d_y}$ *has a* δ-anti-concentrated *distribution if* $\mathbb{P}\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2 = 0\right) \leq \delta$ *for all non-zero* $\boldsymbol{v}$.

In the technical proofs, we are going to apply these properties to terms such as $\left\|\boldsymbol{A}_i^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}^\circ)\right\|_2$ and anti-concentration is essentially asking that any mismatch vector $(\boldsymbol{x} - \boldsymbol{x}^\circ)$ remains "visible" (with some probability) after mapping it through $\boldsymbol{A}_i^\mathsf{T}$; similarly to the proof of Proposition 5, this will allow us to bound $\|\boldsymbol{x} - \boldsymbol{x}^\circ\|_2$ given a bound on $\left\|\boldsymbol{A}_i^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}^\circ)\right\|_2$. Anti-concentration is also related to the probability of sampling a set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, such that $[\boldsymbol{A}_1^\mathsf{T} ; \boldsymbol{A}_2^\mathsf{T} ; \dots ; \boldsymbol{A}_{|\mathcal{I}|}^\mathsf{T}]$ is not full rank. In this sense, it is connected to the notion of non-degeneracy and to the results in Section 6.1 (*e.g.*, if a matrix is not full rank, its smallest singular value is zero, *cf.* Proposition 5).

The sos version of Definition 10 reads as follows.

Definition 11 (Certifiable anti-concentration for matrices, adapted from [37]). *A random matrix* $\boldsymbol{A}$ *has a* k-certifiably (C, δ, M)-anti-concentrated *distribution if there is an even univariate polynomial* p *satisfying* $p(0) = 1$ *such that there is a degree* k *sum-of-squares proof of the following two inequalities:*[15]

$$\forall \boldsymbol{v} , \ \left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2 \leq \delta^2 \quad implies \quad p^2\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2\right) \geq (1 - \delta)^2, \tag{30}$$

$$\forall \boldsymbol{v} , \ \|\boldsymbol{v}\|_2^2 \leq M^2 \quad implies \quad \|\boldsymbol{v}\|^2 \cdot \mathbb{E}\left[p^2\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2\right)\right] \leq C\,\delta\,M^2. \tag{31}$$

A set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, *is* k-certifiably (C, δ, M)-anti-concentrated *if the uniform distribution over the set is* k-certifiably (C, δ, M)-anti-concentrated, *i.e.*,

$$\forall \boldsymbol{v} , \ \left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2^2 \leq \delta^2 \quad implies \quad p^2\left(\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2\right) \geq (1 - \delta)^2, \tag{32}$$

$$\forall \boldsymbol{v} , \ \|\boldsymbol{v}\|_2^2 \leq M^2 \quad implies \quad \|\boldsymbol{v}\|^2 \cdot \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p^2\left(\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2\right) \leq C\,\delta\,M^2. \tag{33}$$

The connection between Definition 10 and its sos counterpart in Definition 11 is not immediate, so a few comments are in order. First, let us consider the probability in Definition 10 and observe that by definition, for any given $\bar{c} \geq 0$, $\mathbb{P}\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2 \leq \bar{c}^2\right) = \mathbb{E}\left[\mathbb{1}\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2 \leq \bar{c}^2\right)\right]$, where $\mathbb{E}\left[\cdot\right]$ is the standard expectation and $\mathbb{1}(\cdot)$ denotes the indicator function, which is such that, for a boolean condition a, $\mathbb{1}(a) = 1$ if the condition is satisfied or zero otherwise. Ideally, we would like to just define certifiable anti-concentration to be satisfied when there is an sos proof for:

$$\mathbb{P}\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2 \leq \bar{c}^2\right) = \mathbb{E}\left[\mathbb{1}\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2 \leq \bar{c}^2\right)\right] \leq \delta. \tag{34}$$

Unfortunately, the indicator function is not a polynomial so such a requirement would not make sense. Therefore, the insight behind Definition 11 is twofold. First, we require the existence of a polynomial p that "behaves" like the indicator function (*cf.* conditions (30) and (32)): this polynomial is required to be close to 1 (*i.e.*, close to the indicator function) whenever the input is smaller than a threshold. Second, we impose such a polynomial to satisfy an anti-concentration bound similar to (34) (*cf.* conditions (31) and (33)). Intuitively, the right-hand-side inequality in (33) prevents that vectors

---

[15]We remark that $p\left(\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2\right)$ in eq. (30) is also a polynomial function of the vector $\boldsymbol{A}^\mathsf{T}\boldsymbol{v}$ since p is an even polynomial, hence it only includes monomials of even degree which can be written as a function of the polynomial $\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2$; the same observation holds for $p\left(\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2\right)$ in eq. (32) with respect to $\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}$.

with large norm $\|\boldsymbol{v}\|$ produce small $\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2$, since that would in turn make $p^2\left(\left\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\right\|_2\right)$ large, hence violating the inequality. Certifiable anti-concentration will allow us to control the norm $\|\boldsymbol{x}-\boldsymbol{x}^\circ\|_2$ from the norm of $\boldsymbol{A}_i^\mathsf{T}(\boldsymbol{x}-\boldsymbol{x}^\circ)$, similarly to what we did in the proofs in Section 6.1.

Certifiable anti-concentration, in its original definition in [37], is a stronger requirement compared to certifiable hypercontractivity, but it still encompasses several relevant distributions, including the standard Gaussian distribution and any anti-concentrated spherically symmetric distributions with strictly sub-exponential tails [37]. In addition, this requirement has been shown to be *necessary* for list-decodable linear regression in [37]. We remark that our definition is slightly different from the one in [37], *e.g.*, we do not normalize the right-hand-side of the inequality $\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2^2 \leq \delta^2$ by the variance of $\left\|\boldsymbol{A}^\mathsf{T}\boldsymbol{v}\right\|_2$ and assume a slightly different implication in (30) and (32); these choices were made for the sake of simplicity, at the cost of some desirable properties of the original definition (*e.g.*, scale invariance) and the need for an additional parameter M in the definition. In Section 8, we empirically show that certifiable anti-concentration is much harder to satisfy in practice (in particular for low-degree polynomials p), but it still applies to some practical problems. We are now ready the present the main results of this monograph.

6.2.1 Estimation Contracts for (LTS)

This section presents estimation contracts for two slightly different estimators based on moment relaxations of the (LTS) problem. The first contract is from [15] (which we adapt to vector-valued measurements) and bounds the residual error of the estimate with respect to the inliers; the second is novel and directly bounds the distance of the estimate with respect to the ground truth.

Algorithm 1: Moment relaxation for (LTS), version 1 [15].

Input: input data $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, inlier rate α, exponent k, relaxation order $r \geq k$.
Output: estimate of $\boldsymbol{x}^\circ$.
/* Algorithm solves a relaxation of the following (LTS) problem: */
/*

$$\min_{\boldsymbol{\omega},\boldsymbol{x},\bar{\boldsymbol{y}}_i,\bar{\boldsymbol{A}}_i, i\in[n]} \left(\frac{1}{n}\sum_{i=1}^n \left\|\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{A}}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2\right)^{\frac{k}{2}} \text{ s.t. } \mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}} \triangleq \begin{cases} \omega_i^2 = \omega_i, \ i \in [n] \\ \sum_{i=1}^n \omega_i = \alpha n \\ \omega_i \cdot (\bar{\boldsymbol{y}}_i - \boldsymbol{y}_i) = \mathbf{0} \ i \in [n] \\ \omega_i \cdot (\bar{\boldsymbol{A}}_i - \boldsymbol{A}_i) = \mathbf{0} \ i \in [n] \\ \boldsymbol{x} \in \mathbb{X} \end{cases} \text{(LTS1)}$$

/* Compute matrix $\boldsymbol{X}^\star$ by solving SDP resulting from moment relaxation */
1 $\boldsymbol{X}^\star = \mathsf{solve\_moment\_relaxation\_at\_order\_}r\,(\text{LTS1})$
/* Pick entries of $\boldsymbol{X}^\star$ corresponding to $\boldsymbol{x}$ */
2 $\boldsymbol{x}_{\mathsf{lts-sdp1}} \triangleq \boldsymbol{X}^\star_{[\boldsymbol{x}]}$
3 **return** $\boldsymbol{x}_{\mathsf{lts-sdp1}}$.

Let us first review the estimator proposed in [15], that we report in Algorithm 1. The algorithm corresponds to Algorithm 5.2 in [15], with the exception that we consider vector-valued measurements and the unknown $\boldsymbol{x}$ in our problem belongs to a basic semi-algebraic set $\mathbb{X}$. The algorithm is based on a moment relaxation of a slightly different (but equivalent) reformulation of (LTS), given in (LTS1) within Algorithm 1. In particular, problem (LTS1) enforces the constraint $\omega_i^2 = \omega_i$, which is equivalent to the constraint $\omega_i \in \{0; 1\}$ in (LTS); moreover, similar to (LTS), it includes the constraints $\sum_{i=1}^n \omega_i = \alpha\,n$ and $\boldsymbol{x} \in \mathbb{X}$. However, problem (LTS1) also includes extra variables

$\bar{\boldsymbol{y}}_i, \bar{\boldsymbol{A}}_i, i \in [n]$, that are such that $\bar{\boldsymbol{y}}_i = \boldsymbol{y}_i$ and $\bar{\boldsymbol{A}}_i = \boldsymbol{A}_i$ whenever $\omega_i = 1$ (which is enforced by the constraints $\omega_i \cdot (\bar{\boldsymbol{y}}_i - \boldsymbol{y}_i) = \boldsymbol{0}, \omega_i \cdot (\bar{\boldsymbol{A}}_i - \boldsymbol{A}_i) = \boldsymbol{0}$), or are zero otherwise (which is enforced by the objective, since whenever $\omega_i = 0$ these variables are no longer constrained). Clearly, this is equivalent to using ω_i in the objective as we did in (LTS). Finally, problem (LTS1) also elevates the objective to the power $k/2$ (where k is an input parameter), which again does not change the optimal solution $\boldsymbol{x}_{\mathsf{LTS}}$ of the non-relaxed problem as compared to the formulation (LTS). In (LTS1), we denoted with $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$ the set of constraints of the problem, since in the technical proofs we will establish sos proofs that connect the set of polynomial constraints $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$ to relevant properties of the estimate $\boldsymbol{x}_{\mathsf{lts-sdp1}}$. After solving the moment relaxation of (LTS1) at order r, Algorithm 1 returns the entries of the (pseudo-moment) matrix $\boldsymbol{X}^\star$ corresponding to $\boldsymbol{x}$, i.e., $\boldsymbol{X}^\star_{[\boldsymbol{x}]}$.

We are now ready to present the estimation contract for Algorithm 1, which is an adaptation of Theorem 5.1 in [15] to vector-valued measurements.[16]

Theorem 12 (Low-outlier case: a priori estimation contract for Algorithm 1, adapted from Theorem 5.1 in [15]). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and known outlier rate β. Call $\ddot{\mathcal{I}}$ the set of uncorrupted measurements $(\boldsymbol{y}_i^\star, \boldsymbol{A}_i^\star)$, $i \in [n]$, where the outliers are replaced by inliers and assume that the set of matrices $\boldsymbol{A}_i^\star$, $i \in \ddot{\mathcal{I}}$, is k-certifiably C-hypercontractive with $k \geq 4$. Moreover, assume $\beta < \beta_{\max} = \sqrt[\frac{k}{2}-1]{1/(C(k/2)^{\frac{k}{2}} 2^{3k-1})}$. Then, Algorithm 1 with relaxation order $r \geq k$ outputs an estimate $\boldsymbol{x}_{\mathsf{lts-sdp1}}$ (not necessarily in $\mathbb{X}$) such that:*

$$\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}_{\mathsf{lts-sdp1}}) \leq (1 + C_1(k,\beta)^{\frac{2}{k}}) \, \mathsf{opt}_{\ddot{\mathcal{I}}} + C_2(k,\beta)^{\frac{2}{k}} \left(\frac{1}{n} \sum_{i=1}^n \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T} \boldsymbol{x}^\star \right\|_2^k \right)^{\frac{2}{k}}, \qquad (35)$$

where $C_1(k,\beta)$ and $C_2(k,\beta)$ are given functions, $\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}) \triangleq \frac{1}{n} \sum_{i=1}^n \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T} \boldsymbol{x} \right\|_2^2$ is the residual error of an estimate $\boldsymbol{x}$ with respect to the inliers $\ddot{\mathcal{I}}$, $\boldsymbol{x}^\star \triangleq \arg\min_{\boldsymbol{x} \in \mathbb{X}} \frac{1}{n} \sum_{i=1}^n \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T} \boldsymbol{x} \right\|_2^2$ is the best estimate from an oracle estimator that has access to all the inliers, and $\mathsf{opt}_{\ddot{\mathcal{I}}} \triangleq \mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}^\star)$ is the corresponding residual error with respect to the inliers $\ddot{\mathcal{I}}$.

The theorem states that —under certifiable hypercontractivity and as long as the outlier rate is below $\beta_{\max}$— the estimate $\boldsymbol{x}_{\mathsf{lts-sdp1}}$ fits well the inlier measurements. In order words, the residual error of $\boldsymbol{x}_{\mathsf{lts-sdp1}}$ over the inliers is not much larger than the best residual $\mathsf{opt}_{\ddot{\mathcal{I}}}$ that an "oracle" estimator that has access to all the inliers would achieve, plus some some higher-order terms that depend on the optimal estimate $\boldsymbol{x}^\star$ returned by the oracle.

The proof of Theorem 12 given in Appendix E is quite involved, but a key tool in it is to bound terms such as $\boldsymbol{A}_i^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}^\star)$ using certifiable hypercontractivity. In particular, the proof derives an sos proof that the constraint set $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$ of (LTS1) implies a desired error bound, and then uses this proof to infer the bound in (35). On the downside, Theorem 12 requires a moment relaxation of order $r \geq k \geq 4$: this requirement stems from the fact that the objective function in (LTS1) is a polynomial of degree $2k$ (which requires a moment relaxation of order at least $r \geq k$) and the fact that the theorem demands $k \geq 4$. Unfortunately, solving moment relaxations of order 4 is currently

---

[16]The reader might notice that, at first glance, the statement Theorem 12 is quite different from the statement of Theorem 5.1 in [15]; however, the substance is identical: we work directly on the set of matrices $\boldsymbol{A}_i$, $i \in [n]$, rather than the generative distribution creating such samples (in the notation of [15], we work on $\widehat{\mathcal{D}}$ instead of $\mathcal{D}$), to avoid the complexity of discussing generalization bounds and requiring bounds on the bit complexity of $\boldsymbol{x}^\star$, which do not add much to the discussion in this monograph.

impractical.[17] For instance, related work [30] relies on relaxations of order 2 and still has to develop sparse approximations and ad-hoc solvers to make them run in a reasonable time. Moreover, the bound (35) does not immediately inform us about how far the estimate is from the ground truth.

In the rest of this section we present a second algorithm and the corresponding estimation contract that directly quantifies the distance between the estimate and the ground truth. This result is novel, but relies on two lemmas proposed for a different goal in [37].

Let us start by stating a slightly different (LTS)-like estimator in Algorithm 2.

Algorithm 2: Moment relaxation for (LTS), version 2.

Input: input data $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, inlier rate α, relaxation order $r \geq 2$.
Output: estimate of $\boldsymbol{x}^{\circ}$.
/* Algorithm solves a relaxation of the following (LTS)-like problem: */
/*

$$\min_{\boldsymbol{\omega}, \boldsymbol{x}} \ \frac{1}{n} \sum_{i=1}^n \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x} \right\|_2^2 \ \texttt{s.t.} \ \mathcal{T}_{\boldsymbol{\omega}, \boldsymbol{x}} \triangleq \left\{ \begin{array}{c} \omega_i^2 = \omega_i, \ \ i \in [n] \\ \sum_{i=1}^n \omega_i = \alpha n \\ \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x} \right\|_2^2 \leq \bar{c}^2 \ \ i \in [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\} \quad \text{(LTS2)}$$

/* Compute matrix $\boldsymbol{X}^{\star}$ by solving SDP resulting from moment relaxation */
1 $\boldsymbol{X}^{\star} = \texttt{solve\_moment\_relaxation\_at\_order\_}r\,(\text{LTS2})$
/* Compute estimate */

2 for each $i \in [n]$ set: $\boldsymbol{v}_i = \left\{ \begin{array}{ll} \dfrac{\boldsymbol{X}^{\star}_{[\omega_i \boldsymbol{x}]}}{\boldsymbol{X}^{\star}_{[\omega_i]}} & \text{if } \boldsymbol{X}^{\star}_{[\omega_i]} > 0 \\ \boldsymbol{0} & \text{otherwise} \end{array} \right.$

3 $\boldsymbol{x}_{\mathsf{lts-sdp2}} = \sum_{i=1}^n \dfrac{\boldsymbol{X}^{\star}_{[\omega_i]}}{\sum_{j=1}^n \boldsymbol{X}^{\star}_{[\omega_j]}} \boldsymbol{v}_i$

4 **return** $\boldsymbol{x}_{\mathsf{lts-sdp2}}$.

Algorithm 2 is fairly different from Algorithm 1. First of all, it is based on a relaxation of an "(LTS)-like" problem: the non-relaxed problem (LTS2) has additional constraints $\omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x} \right\|_2^2 \leq \bar{c}^2$ that do not appear in (LTS), but it can be seen to be equivalent to (LTS) with $f_i(\boldsymbol{x}) = \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}$ otherwise.[18] Moreover, while Algorithm 2 also uses a moment relaxation, it computes an estimate $\boldsymbol{x}_{\mathsf{lts-sdp2}}$ by averaging multiple vectors extracted from the solution of the moment relaxation (lines 2-3).

We provide the following estimation contract for Algorithm 2.

---

[17]Recall from Section 4 that the moment matrix has size $\binom{d+r}{r} \times \binom{d+r}{r}$ where d is the dimension of the variables in the polynomial optimization problem (LTS1) and r is the relaxation order; note that the dimension d grows with the number of measurements n, and in practical problems $n \gg 100$. Therefore, the matrix quickly becomes too large to handle for current SDP solvers. As an example, for $d = 100$ and $r = 4$, $\binom{d+r}{r} = 4\,598\,126$.

[18]Note that the pair $(\boldsymbol{x}^{\circ}, \boldsymbol{\omega}^{\circ})$, where $\omega_i^{\circ} = 1$ if $i \in \mathcal{I}$ and zero otherwise, satisfies all the constraints in (LTS2), hence the problem is feasible.

> **Proposition 13** (Low-outlier case: a priori estimation contract for Algorithm 2). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and outlier rate $\beta < 0.5$ (or, equivalently, inlier rate $\alpha = 1 - \beta > 0.5$). Call $\mathcal{I}$ the set of inliers and assume that the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2 \eta^2 (1-2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated for some $\eta > 0$. Then, Algorithm 2 with relaxation order $r \geq k/2$ outputs an estimate $\boldsymbol{x}_{\mathsf{lts-sdp2}}$ (not necessarily in $\mathbb{X}$) such that:*
>
> $$\|\boldsymbol{x}_{\mathsf{lts-sdp2}} - \boldsymbol{x}^\circ\|_2 \leq M_x \left(\frac{\alpha\, \eta}{2} + 2\frac{1-\alpha}{\alpha} \right). \tag{36}$$

While Proposition 13 takes a stronger assumption on the input data (see the discussion on certifiable hypercontractivity vs. certifiable anti-concentration in Section 6.2), it provides a direct bound on the distance of the estimate from the ground truth $\boldsymbol{x}^\circ$. Fig. 2 plots the bound on the right-hand-side of eq. (36) as a solid blue line, for $M_x = 1$, $\eta = 0.01$, and inlier rates α between 0 and 1. For comparison, we also report the trivial bound $\|\boldsymbol{x}_{\mathsf{lts-sdp2}} - \boldsymbol{x}^\circ\|_2 \leq 2M_x$ as a dashed red line.[19] We observe that, as expected, the proposed bound is only informative in the low-outlier case (*i.e.*, when $\alpha > 0.5$). Moreover, the bound predicts decreasing estimation errors when the number of inliers increases, consistently with our expectations, and approaches $\frac{M_x \eta}{2}$ as α approaches 1.
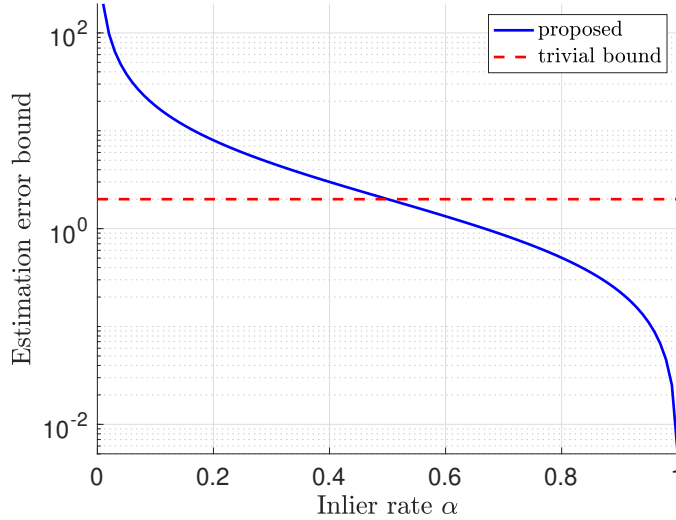
Figure 2: Comparison between the proposed bound in eq. (36) with $M_x = 1$, $\eta = 0.01$ (solid blue line) and the trivial bound $\|\boldsymbol{x}_{\mathsf{lts-sdp2}} - \boldsymbol{x}^\circ\|_2 \leq 2M_x$ (dashed red line) for inlier rates $\alpha \in [0, 1]$.

We conclude this section by noting that —while in principle Algorithm 2 might require a relaxation of order as low as 2, similar to the ones commonly used in practice (see [30])— the choice of the order is still dictated by the anti-concentration properties of the input data, due to the constraint $r \geq k/2$, where r is the order of the moment relaxation, while k depends on the anti-concentration properties of the inliers. As we will see in Section 8, k will need to be ≥ 6 (due to the degree of the polynomials involved in (33)), thus again restricting the practicality of the result in Proposition 13, which in practice would only be applicable to relaxations of order ≥ 3.

---

[19]The bound follows from the triangle inequality $\|\boldsymbol{x}_{\mathsf{lts-sdp2}} - \boldsymbol{x}^\circ\|_2 \leq \|\boldsymbol{x}_{\mathsf{lts-sdp2}}\|_2 + \|\boldsymbol{x}^\circ\|_2 \leq 2M_x$.

6.2.2 Estimation Contract for (MC)

This section presents an estimation contract for Algorithm 3 below, which is based on a moment relaxation of the (MC) problem. Notice that (MC1) in Algorithm 3 is equivalent to (MC) for $f_i(\boldsymbol{x}) = \boldsymbol{A}_i^\top \boldsymbol{x}$. Let us start by presenting the algorithm the estimation contract refers to.

Algorithm 3: Moment relaxation for (MC).

Input: input data $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, relaxation order $r \geq 2$.

Output: estimate of $\boldsymbol{x}^\circ$.

/* Algorithm solves a relaxation of the following (MC) problem: */

/*

$$
\max_{\boldsymbol{\omega}, \boldsymbol{x}} \sum_{i=1}^{n} \omega_i, \qquad \texttt{s.t.} \ \ \mathcal{M}_{\boldsymbol{\omega}, \boldsymbol{x}} \triangleq \left\{ \begin{array}{c} \omega_i^2 = \omega_i, \ \ i \in [n] \\ \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\top \boldsymbol{x} \right\|_2^2 \leq \bar{c}^2 \ \ i \in [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\} \qquad \text{(MC1)}
$$

/* Compute matrix $\boldsymbol{X}^\star$ by solving SDP resulting from moment relaxation   */

1 $\boldsymbol{X}^\star = \texttt{solve\_moment\_relaxation\_at\_order\_}r\,(\text{MC1})$

/* Compute estimate   */

2 for each $i \in [n]$ set: $\boldsymbol{v}_i = \begin{cases} \dfrac{\boldsymbol{X}^\star_{[\omega_i \boldsymbol{x}]}}{\boldsymbol{X}^\star_{[\omega_i]}} & \text{if } \boldsymbol{X}^\star_{[\omega_i]} > 0 \\ \boldsymbol{0} & \text{otherwise} \end{cases}$

3 $\boldsymbol{x}_{\mathsf{mc-sdp}} = \sum_{i=1}^{n} \dfrac{\boldsymbol{X}^\star_{[\omega_i]}}{\sum_{j=1}^{n} \boldsymbol{X}^\star_{[\omega_j]}} \boldsymbol{v}_i$

4 return $\boldsymbol{x}_{\mathsf{mc-sdp}}$.

Contrary to the algorithms in Section 6.2.1, Algorithm 3 does not require prior knowledge about the number of inliers $\alpha\, n$. The moment relaxation in Algorithm 3 is similar to the one proposed in [30];[20] the computation of the estimate $\boldsymbol{x}_{\mathsf{mc-sdp}}$ is different from the one in [30], but the two "rounding" schemes are equivalent when the relaxation is tight (in which case $\boldsymbol{X}^\star_{[\omega_i]} \in \{0, 1\}$ and $\boldsymbol{X}^\star_{[\omega_i \boldsymbol{x}]} = \boldsymbol{X}^\star_{[\boldsymbol{x}]}$ $\forall i$ such that $\boldsymbol{X}^\star_{[\omega_i]} > 0$). We provide the following estimation contract for Algorithm 3.

Proposition 14 (Low-outlier case: a priori estimation contract for Algorithm 3)**.** *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and outlier rate $\beta < 0.5$ (or, equivalently, inlier rate $\alpha = 1 - \beta > 0.5$). Call $\mathcal{I}$ the set of inliers and assume that the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2 \eta^2 (1 - 2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated for some $\eta > 0$. Then, Algorithm 3 with relaxation order $r \geq k/2$ outputs an estimate $\boldsymbol{x}_{\mathsf{mc-sdp}}$ (not necessarily in $\mathbb{X}$) such that:*

$$
\left\| \boldsymbol{x}_{\mathsf{mc-sdp}} - \boldsymbol{x}^\circ \right\|_2 \leq M_x \left(\frac{\alpha\, \eta}{2} + 2\frac{1 - \alpha}{\alpha} \right). \tag{37}
$$

Note that the performance guarantees are essentially the same as Proposition 13 and the proof proceeds along the same line, but in this case the objective forces the solution to pick enough inliers, while in Proposition 13 the same role was played by the constraint $\sum_{i=1}^{n} \omega_i = \alpha\, n$.

---

[20]The presentation in [30] uses the slightly different but equivalent parametrization where the binary variables are restricted to $\omega_i \in \{-1; +1\}$ and their binary nature is enforced via $\omega_i^2 = 1$, while here we use a more straightforward parametrization with $\omega_i \in \{0; 1\}$ and enforce it via $\omega_i^2 = \omega_i$.

6.2.3 Estimation Contract for (TLS)

This section presents an estimation contract for Algorithm 4, which is based on a moment relaxation of the (TLS) problem. Notice that (TLS1) in Algorithm 4 is equivalent to (TLS) for $f_i(\boldsymbol{x}) = \boldsymbol{A}_i^\mathsf{T}\boldsymbol{x}$.[21]

Algorithm 4: Moment relaxation for (TLS).

Input: input data $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, relaxation order r.
Output: estimate of $\boldsymbol{x}^\circ$.
/* Algorithm solves a relaxation of the following (TLS) problem: */
/*

$$\min_{\boldsymbol{x},\boldsymbol{\omega}} \quad \sum_{i=1}^{n} \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T}\boldsymbol{x} \right\|_2^2 + (1 - \omega_i) \cdot \bar{c}^2 \tag{TLS1}$$

$$\texttt{s.t.} \quad \mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \triangleq \left\{ \begin{array}{c} \omega_i^2 = \omega_i, \ \ i \in [n] \\ \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T}\boldsymbol{x} \right\|_2^2 \le \bar{c}^2 \ \ i \in [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\}$$

/* Compute matrix $\boldsymbol{X}^\star$ by solving SDP resulting from moment relaxation   */
1 $\boldsymbol{X}^\star = \mathsf{solve\_moment\_relaxation\_at\_order\_}r$ (TLS1)
/* Compute estimate   */
2 for each $i \in [n]$ set: $\boldsymbol{v}_i = \begin{cases} \dfrac{\boldsymbol{X}^\star_{[\omega_i \boldsymbol{x}]}}{\boldsymbol{X}^\star_{[\omega_i]}} & \text{if } \boldsymbol{X}^\star_{[\omega_i]} > 0 \\ \boldsymbol{0} & \text{otherwise} \end{cases}$
3 $\boldsymbol{x}_{\mathsf{tls-sdp}} = \sum_{i=1}^{n} \dfrac{\boldsymbol{X}^\star_{[\omega_i]}}{\sum_{j=1}^{n} \boldsymbol{X}^\star_{[\omega_j]}} \boldsymbol{v}_i$
4 **return** $\boldsymbol{x}_{\mathsf{tls-sdp}}$.

Similar to Algorithm 3 in Section 6.2.2, Algorithm 4 does not require prior knowledge about the number of inliers αn. Moreover, the moment relaxation is similar to the one proposed in [30] (except from the redundant constraints $\omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T}\boldsymbol{x} \right\|_2^2 \le \bar{c}^2$ in $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$); the computation of the estimate $\boldsymbol{x}_{\mathsf{tls-sdp}}$ is different from the one in [30], but the two "rounding" schemes are equivalent when the relaxation is tight. We provide the following estimation contract for Algorithm 4.

---

[21] The attentive reader might notice that $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$ in (TLS1) includes redundant constraints $\omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T}\boldsymbol{x} \right\|_2^2 \le \bar{c}^2$ in $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$: these constraints do not change the optimal solution of (TLS) —since the objective is already forcing $\omega_i = 0$ whenever $\left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T}\boldsymbol{x} \right\|_2^2 > \bar{c}^2$— but will make our proofs easier.

Proposition 15 (Low-outlier case: a priori estimation contract for Algorithm 4). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and outlier rate $\beta < 0.5$ (or, equivalently, inlier rate $\alpha = 1 - \beta > 0.5$). Call $\mathcal{I}$ the set of inliers and assume that the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2 \eta^2 (1-2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated for some $\eta > 0$. Then, Algorithm 4 with relaxation order $r \geq k/2$ outputs an estimate $\boldsymbol{x}_{\mathsf{tls-sdp}}$ (not necessarily in $\mathbb{X}$) such that:*

$$\|\boldsymbol{x}_{\mathsf{tls-sdp}} - \boldsymbol{x}^\circ\|_2 \leq \frac{\alpha M_x n}{\alpha\, n - \frac{\gamma^\circ}{\bar{c}^2}} \left(\frac{\alpha\, \eta}{2} + 2\frac{1-\alpha}{\alpha} \right), \tag{38}$$

where $\gamma^\circ \triangleq \sum_{i \in \mathcal{I}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ \right\|_2^2$ is the squared residual error of the ground truth $\boldsymbol{x}^\circ$ over the inliers $\mathcal{I}$.

Contrarily to Proposition 13 and Proposition 14, the error bound in Proposition 15 depends on the ground truth residual error γ°. In particular, if $\gamma^\circ = 0$ (*i.e.,* noiseless inliers), the bound becomes the same as the ones in Proposition 13 and Proposition 14, but as the residual error approaches $\gamma^\circ = \alpha\, n\, \bar{c}^2$ (*i.e.,* each inlier has the maximum allowed error $\bar{c}^2$), the bound become vacuous.

7 Estimation Contracts for High Outlier Rates

This section focuses on the high-outlier case where $\beta \gg 0.5$. Note that in the high-outlier case any point estimator —including (LTS), (MC), and (TLS)— can be tricked into returning an arbitrarily wrong estimate: intuitively, since the majority of the measurements are outliers, the outliers can agree on an $\boldsymbol{x}$ that optimizes (LTS), (MC), or (TLS), while being far from $\boldsymbol{x}^\circ$. At the same time, an algorithm can still compute a provably accurate estimate if it is allowed to return a *list* of potential hypotheses, in the hope that at least one of them is correct. This setup is typically referred to as *list-decodable regression* [37]. In the following, we review the algorithm and theoretical guarantees from [37], which we adapt to the case of vector-valued measurements.

Remark 16 (High vs. Low Outlier Rates). *While the algorithms and results in this section are designed for the high-outlier case (i.e., $\beta > 0.5$) they are also applicable to the low-outlier case; in other words, the estimation contracts in this section will not require $\beta > 0.5$. At the same time, if one can assume that $\beta \ll 0.5$, it remains convenient to use the algorithms in Section 6 since those return a unique estimate, which is typically desirable in practice.*

7.1 A Priori Estimation Contract for List-Decodable Estimation

In the high-outlier case, Karmalkar *et al.* [37] proposed using Algorithm 5 (given below) to perform list-decodable outlier-robust regression. The algorithm returns a (small) list of potential estimates, and, as we will see below, under suitable conditions on the inliers, it guarantees that at least one of such estimates is close to the ground truth. Both the algorithm and estimation contract in this section are adaptations of the results in [37] to the case of vector-valued measurements.

The rationale behind problem (LDR) is not as immediate as the problems presented in the previous sections. For instance, in (MC1), it was clear we tried to select the largest number of measurements, *i.e.,* the largest $\sum_{i=1}^n \omega_i$, such that the corresponding estimate $\boldsymbol{x}$ had residual below $\bar{c}$. However, (LDR) seems to do just the opposite: it is looking for the *smallest* $\sum_{i=1}^n \omega_i^2$.

Intuitions behind Algorithm 5. Imagine we could build a list $\hat{\tilde{\mathcal{S}}}$ containing every possible subset $\mathcal{S}'$ of $\alpha\, n$ measurements that satisfy $\left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}' \right\|_2^2 \leq \bar{c}^2, \forall i \in \mathcal{S}'$ and for some estimate $\boldsymbol{x}'$. Clearly, the set of inliers $\mathcal{I}$ would be one of such subsets and hence belong to $\hat{\tilde{\mathcal{S}}}$. For each such

---
Algorithm 5: List-Decodable Outlier-Robust Estimation [37].

> **Input:** input data $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, inlier rate α, relaxation order $r \geq 2$, integer $N \geq 1$.
> **Output:** list of estimates of $\boldsymbol{x}^\circ$.
> /* Algorithm solves a relaxation of the following problem: */
> /*
>
> $$\min_{\boldsymbol{\omega},\boldsymbol{x}} \|\boldsymbol{\omega}\|_2^2, \ \ \texttt{s.t.} \ \ \mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}} \triangleq \left\{ \begin{array}{c} \omega_i^2 = \omega_i, \ \ i = [n] \\ \sum_{i=1}^n \omega_i = \alpha n \\ \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x} \right\|_2^2 \leq \bar{c}^2, \ \ i = [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\} \quad \text{(LDR)}$$
>
> /* Compute matrix $\boldsymbol{X}^\star$ by solving SDP resulting from moment relaxation   */
> 1 $\boldsymbol{X}^\star = \mathsf{solve\_moment\_relaxation\_at\_order}\_r \, (\text{LDR})$
> /* Compute list of estimates   */
> 2   for each $i \in [n]$ set: $\boldsymbol{v}_i = \left\{ \begin{array}{ll} \dfrac{\boldsymbol{X}^\star_{[\omega_i \boldsymbol{x}]}}{\boldsymbol{X}^\star_{[\omega_i]}} & \text{if } \boldsymbol{X}^\star_{[\omega_i]} > 0 \\ \boldsymbol{0} & \text{otherwise} \end{array} \right.$
> 3   create empty list $\mathcal{L} = \emptyset$.
> 4   sample N/α times from $[n]$ with probability $\frac{1}{\alpha n} \boldsymbol{X}^\star_{[\omega_i]}$, and for each extracted i add $\boldsymbol{v}_i$ to $\mathcal{L}$.
> 5   **return** $\mathcal{L}$.
---

subset $\mathcal{S}' \in \hat{\hat{\mathcal{S}}}$, we could define the indicator vector $\boldsymbol{\omega}' = \mathbf{1}_{\mathcal{S}'}$, and by inspection, $(\boldsymbol{\omega}', \boldsymbol{x}')$ would be feasible for (LDR). Therefore, the first intuition behind Algorithm 5 is that the feasible set of $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ of (LDR) includes every such $(\boldsymbol{\omega}', \boldsymbol{x}')$, including $(\boldsymbol{\omega}_{\mathcal{I}}, \boldsymbol{x}^\circ)$. Unfortunately, the list $\hat{\hat{\mathcal{S}}}$ has exponential size in general [37]. Therefore, the second intuition behind Algorithm 5 is that the moment relaxation (LDR) will compute an indicator vector $\boldsymbol{\omega}$ that will "spread" across many subsets in $\hat{\hat{\mathcal{S}}}$. While this is formalized in the proofs (and further discussed in [37]), it is relatively easy to visualize why the relaxation would "spread" across the subsets by considering the following simpler relaxation of (LDR):

$$\min_{\boldsymbol{\omega},\boldsymbol{x}} \|\boldsymbol{\omega}\|_2^2, \ \ \text{s.t.} \ \ \left\{ \begin{array}{c} \omega_i \in [0,1], \ \ i = [n] \\ \sum_{i=1}^n \omega_i = \alpha \, n \\ \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x} \right\|_2^2 \leq \bar{c}^2, \ \ i = [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\} \quad (39)$$

where we only relaxed $\omega_i \in \{0; 1\}$ to $\omega_i \in [0, 1]$. Imagine that the list $\hat{\hat{\mathcal{S}}}$ contains two overlapping sets $\mathcal{S}'$ and $\mathcal{S}''$ and hence $\boldsymbol{\omega}' = \mathbf{1}_{\mathcal{S}'}$ and $\boldsymbol{\omega}'' = \mathbf{1}_{\mathcal{S}''}$ would be feasible for (39) for some $\boldsymbol{x}$. In such a case the vector $\boldsymbol{\omega} = \frac{\mathbf{1}_{\mathcal{S}'} + \mathbf{1}_{\mathcal{S}''}}{2}$ —that spreads across both subsets— would still satisfy the constraint $\sum_{i=1}^n \omega_i = \alpha \, n$ and achieve a lower objective in (39) compared to $\mathbf{1}_{\mathcal{S}'}$ and $\mathbf{1}_{\mathcal{S}''}$.[22] Similarly, after applying the moment relaxation to (LDR), the objective of (LDR) will favor indicator vectors that span multiple sets. The last intuition behind Algorithm 5 is that we can use these indicator vectors to sample good estimates of $\boldsymbol{x}^\circ$, as done in lines 2-4 of the algorithm.

    Estimation contract. The estimation contract for Algorithm 5 is given as follows.

---
[22]Calling t the number of common elements between $\mathcal{S}'$ and $\mathcal{S}''$: $\|\boldsymbol{\omega}\|_2^2 = \frac{1}{4}(\|\mathbf{1}_{\mathcal{S}'} + \mathbf{1}_{\mathcal{S}''}\|_2^2) = \frac{1}{4}(2\alpha \, n + 2t) = \frac{\alpha \, n + t}{2}$ (which is $< \alpha \, n$ as long as the sets do not fully overlap), while each set would achieve an objective $\|\boldsymbol{\omega}'\|_2^2 = \|\boldsymbol{\omega}''\|_2^2 = \alpha \, n$.

Theorem 17 (High-outlier case: a priori estimation contract for Algorithm 5, adapted from Theorem 1.5 in [37]). *Consider Problem 1 with measurements* $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, *and known outlier rate* β *(or, equivalently, known inlier rate* $\alpha = 1 - \beta$), *possibly with* $\beta > 0.5$. *Call* $\mathcal{I}$ *the set of inliers and assume that the set of matrices* $\boldsymbol{A}_i$, $i \in \mathcal{I}$, *is* k-*certifiably* $(\frac{\alpha^2 \eta^2 (1 - 2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-*anti-concentrated for some* $\eta > 0$. *Then, with probability at least* $1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{N}{\alpha}}$ *(over the draw of the samples in the algorithms), where* $N \geq 1$ *is a user-defined parameter, Algorithm 5 with relaxation order* $r \geq k/2$ *outputs a list* $\mathcal{L}$ *of size* N/α *such that there is an estimate* $\boldsymbol{x} \in \mathcal{L}$ *(with* $\boldsymbol{x}$ *not necessarily in* $\mathbb{X}$*) such that*

$$\|\boldsymbol{x} - \boldsymbol{x}^\circ\|_2 \leq \eta M_x. \tag{40}$$

Moreover, when $\alpha \geq 0.01$ *(i.e., at least 1% of the measurements are inliers) and* $N = 10$, *the relation* $\|\boldsymbol{x} - \boldsymbol{x}^\circ\|_2 \leq \eta M_x$ *holds with probability at least* 0.99 *over the draw of the samples.*

The estimation contract in Theorem 17 provides probabilistic guarantees on the outcome of the estimator, as opposed to the deterministic results in the previous sections. The sampling stage is needed to keep the list small, while also ensuring that it captures at least an estimate close to the ground truth. Note that the final statement in the theorem is just a particularization of the claim that eq. (40) holds with probability at least $1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{N}{\alpha}}$; this is given to reassure the reader that the approach can provide good estimates with high probability already for small N (*i.e.*, $N = 10$) and for very challenging problems where α is very low (*i.e.*, $\alpha = 0.01$). Clearly, the probability will increase with the user-specified parameter N, which controls the number of samples to draw.

Theorem 17 and Algorithm 5 are of interest for geometric perception for several reasons. First of all, contrary to the literature on multi-hypothesis estimation (*e.g.*, [42, 44, 45]) and particle filtering (*e.g.*, [43]) in robotics and related fields, Algorithm 5 provides performance guarantees while retaining a small number of hypotheses; on the other hand, in multi-hypothesis tracking, one either has to retain an exponential number of hypotheses or typically loses performance guarantees, while in particle filtering one typically only has asymptotic guarantees for growing number of samples. Furthermore, while the sampling scheme might carry some resemblance of RANSAC [51], the number of samples (*i.e.*, iterations) required by RANSAC grows exponentially with the size of the minimal set (which, in turn, depends on the dimension of $\boldsymbol{x}^\circ$) and the outlier rate [62]; this is in stark contrast with Algorithm 5, where the number of samples is independent on the dimension of $\boldsymbol{x}^\circ$ and only grows as N/α (also note that the number of samples in Algorithm 5 is upper bounded by n).

In the following section we show that while the guarantees in Theorem 17 fall short in practical problems, a simple modification of Algorithm 5 with $r = 2$ has impressive performance in practice.

8 Numerical Experiments

This section elucidates on the theoretical results, by (i) assessing the tightness and the applicability of the error bounds discussed in this monograph, (ii) evaluating how strict are the assumptions made on the input data (*i.e.*, certifiable hypercontractivity or certifiable anti-concentration), and (iii) assessing the practical performance of (a variant of) Algorithm 5 for list-decodable estimation. The analysis is performed on a canonical perception problem, namely *rotation search* (see Example 1), which finds application in satellite attitude estimation [98] and image stitching [34] problems, among others. After introducing the experimental setup in Section 8.1, Section 8.2 focuses on the a posteriori bound introduced in Proposition 6, Section 8.3 analyzes the certifiable hypercontractivity assumption and the contract in Theorem 12, Section 8.4 analyzes the certifiable anti-concentration assumption and

the contract in Proposition 6. Section 8.5 concludes the experiments with an empirical evaluation of list-decodable estimation.

8.1 Experimental Setup

We use *rotation search* (also known as the *Wahba problem*) [98, 34] as a prototypical outlier-robust estimation problem, see Example 1. The goal of the problem is to estimate a rotation $\boldsymbol{R}$ that aligns two sets of 3D vectors (typically normalized to have unit norm). The inputs to the problem are the vector pairs $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{R}^3$, $i = 1, \ldots, n$. The vectors $\boldsymbol{a}_i$ are used to build the measurement matrix $\boldsymbol{A}_i$ as in Example 1; these vectors are either known or measured: for instance, in satellite pose estimation these vectors are read from a star database, while in image stitching they correspond to detected image features. $\boldsymbol{b}_i$ are measured bearing vectors. The association between pairs of vectors $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ is obtained using image processing or learning-based methods and is prone to errors, making a large number of measurements outliers. We test on both synthetic and real data, as described below.

Synthetic data. We use the same data generation protocol of [34]. In particular, we create each test instance by sampling n vectors $\{\boldsymbol{a}_i\}_{i=1}^n$ uniformly at random on the unit sphere. Then, we pick a random rotation $\boldsymbol{R}$, and apply it to each $\boldsymbol{a}_i$ according to eq. (4) and add zero-mean Gaussian noise with covariance $10^{-4} \cdot \mathbf{I}_3$ to get $\{\boldsymbol{b}_i\}_{i=1}^n$. To generate outliers, we replace a fraction β of $\boldsymbol{b}_i$'s with random unit-norm vectors. We set the maximum error bound $\bar{c}^2$ from the quantile of the χ^2 distribution with three degrees of freedom and lower tail probability equal to 0.9999 (see Remark 1 in [34]), thus obtaining $\bar{c}^2 = 0.0021$ (which also accounts for the measurement variance). Results on synthetic data are averaged over 100 Monte Carlo runs unless specified otherwise.

(a)   (b)
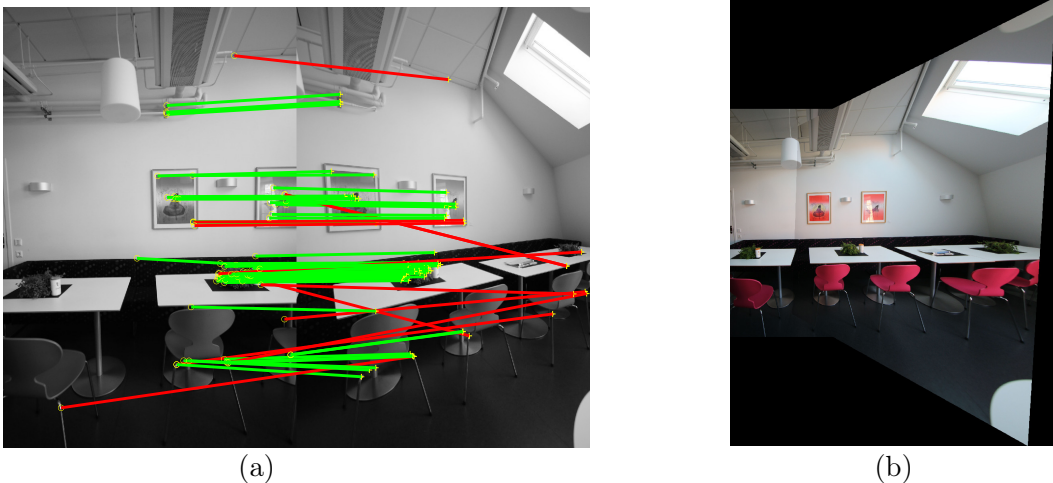
Figure 3: Image stitching example: (a) Feature matches between two images from which the vectors $(\boldsymbol{a}_i, \boldsymbol{b}_i)$, $i = 1, \ldots, n$, are generated; the matches include many outliers, shown in red (inliers are shown in green). (b) Image stitching results using the moment relaxation approach in [34].

Real data. We use the PASSTA dataset to test rotation search problems arising in image stitching applications [117] and use the same data generation protocol of [34]. In particular, to generate the vector pairs $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{R}^3$, $i = 1, \ldots, n$, we first use SURF [118] to detect and match point features between the two images. Fig. 3(a) shows the matches for one of the image pairs in the dataset. From the SURF feature points, we apply the inverse of the known camera intrinsic matrix $\boldsymbol{K}$ to obtain unit-norm bearing vectors $\{\boldsymbol{a}_i, \boldsymbol{b}_i\}_{i=1}^{70}$ observed in each camera frame. Then, we solve
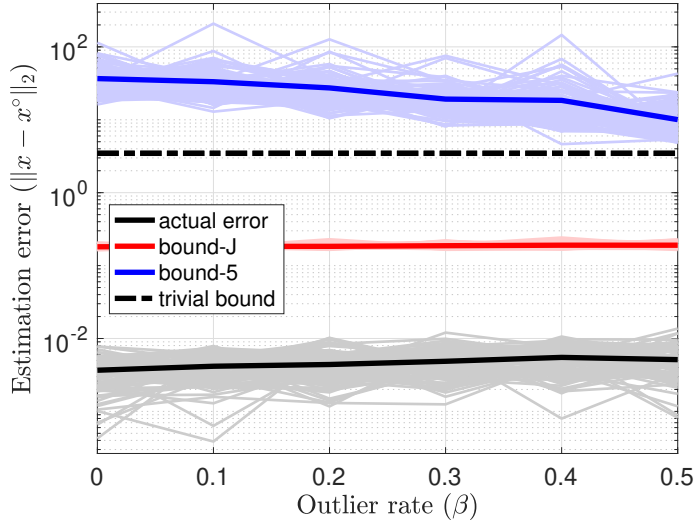
Figure 4: Actual error of the (TLS) estimator compared to the a posteriori bounds from Proposition 6 and the trivial bound $2M_x = 2\sqrt{3}$, with $n = 50$ and increasing outlier rates β. We report a posteriori bounds for $d_{\mathcal{J}} = 5$ ("bound-5") and for the set $\mathcal{J}$ chosen as the correctly selected inliers ("bound-J").

the rotation search problem to estimate the rotation $\boldsymbol{R}$ between the two camera frames, using the outlier-corrupted pairs $\{\boldsymbol{a}_i, \boldsymbol{b}_i\}_{i=1}^{70}$. Finally, using the estimated $\boldsymbol{R}$, we can compute the homography matrix as $\boldsymbol{H} = \boldsymbol{KRK}^{-1}$ to stitch the pair of images together; see the example in Fig. 3(b).

Algorithms and implementation details. We use the sparse moment relaxation proposed in [34] to solve the (TLS) formulation of the rotation search problem, and use STRIDE [30, 119] as an SDP solver for the resulting semidefinite relaxation. Note that the SDP proposed in [34] is a sparse version of the relaxation in Algorithm 4 (with order $r = 2$), and the output of the two relaxations match whenever the sparse relaxation is tight; in all results below we observe the sparse moment relaxation of (TLS) to be tight (*i.e.,* relaxation gap below 10^{-7}) hence we do not differentiate between the two algorithms. We use cvx [106] and MOSEK [105] as a parser/solver for the SDPs arising in list-decodable estimation. The numerical results are obtained on a MacBook Pro with 2.8 GHz Quad-Core Intel Core i7 processor. The Matlab code to reproduce the experiments below can be found at https://github.com/MIT-SPARK/estimation-contracts.

8.2 Low Outlier Rates: A Posteriori Bounds

This section evaluates the quality of the a posteriori bound in Proposition 6. The bound applies since in all the tests in this section the relaxation from [34] is tight and hence it obtains an optimal solution to (TLS). Moreover, with our data generation method, every subset of $\bar{d} = 3$ inliers is nondegenerate by construction.[23] We note that the bound holds for any choice of integer $d_{\mathcal{J}} = \bar{d}, \ldots, (2\alpha - 1)n - \frac{\gamma^\circ}{\bar{c}^2}$. For a small $d_{\mathcal{J}}$, we can compute the bound by brute-force testing every subset $\mathcal{J}$ of the set of inliers computed by (TLS) and exhaustively searching the smallest value of $\sigma_{\min}(\boldsymbol{A}_{\mathcal{J}})$; in our tests, we are able to compute the bound for $d_{\mathcal{J}} = 5$ in a matter of seconds. However, we expect large values of $d_{\mathcal{J}}$ to produce more informative bounds (intuitively, in the proof of Proposition 6 those bounds leverage

---

[23]In rotation search, the unknown rotation can be uniquely estimated from $\bar{d} = 3$ non-collinear 3D points, and our data generation method for the 3D vectors $\boldsymbol{a}_i$ produces triplets of collinear points with probability zero.

a larger number of measurements); since the bound Proposition 6 is expensive to compute for large $d_{\mathcal{J}}$, we manually compute the set $\mathcal{J}$ corresponding to the intersection of the measurements selected by (TLS) as inliers and the true inliers, and compute the resulting bound for that choice of $\mathcal{J}$: this is the set used in our proofs to derive the estimation error bounds; note that such a set $\mathcal{J}$ can only be computed in simulation, since in practice the true inliers are unknown.

Figure 4 compares the actual estimation error with the bounds from Proposition 6. The estimation error is computed as $\|\boldsymbol{x}_{\mathsf{TLS}} - \boldsymbol{x}^\circ\|_2$, where $\boldsymbol{x}$ and $\boldsymbol{x}^\circ$ are the vectorized representations of the estimated and the ground-truth rotation, respectively. For the bounds, we plot the bound we compute for $d_{\mathcal{J}} = 5$ (label: "bound-5") and the bound obtained from the choice of $\mathcal{J}$ described above (label: "bound-J"). We also plot the trivial bound $2M_x = 2\sqrt{3}$ computed by using the triangle inequality on $\|\boldsymbol{x}_{\mathsf{TLS}} - \boldsymbol{x}^\circ\|_2$ and by noting that every (vectorized) rotation has norm $\sqrt{3}$. Several comments are in order. First, the actual error increases with the outlier rate (the trend is slightly more difficult to see due to the log scale): this is expected since with increasing outlier rates the number of "useful" measurements decreases; however, the error remains very small (*i.e.,* less than 1 degree) for all outlier rates. This confirms the empirical robustness of moment relaxations of (TLS) already observed in related work [34, 30], which have been reported to achieve impressive performance even for much higher rates of random outliers. Second, bound-5 is unfortunately too loose and uninformative, since —while it improves for larger outlier rates[24]— it remains larger than the trivial bound. On the other hand, bound-J is significantly better than the trivial bound. While there is still a large gap between the actual error and bound-J, we remark that the bound considers the worst-case scenario where the outliers are possibly adversarial; on the other hand, the outlier generation mechanism we borrow from [34] is relatively benign, since the outliers are sampled independently, at random, and without knowledge of the inliers. In other words, there might well be instances where the actual error could be much closer to bound-J.

8.3 Low Outlier Rates: A Priori Bounds and Hypercontractivity

This section shows that k-certifiable hypercontractivity is a relatively mild assumption for $k = 4$ and for choices of $C(t)^t$ as small as 5 or 6. This observation is validated on both synthetic and real data. On the other hand, the experiments offer a closer look at the bound in Theorem 12, which in hindsight only applies to relatively "easy" problems with very small amounts of outliers.

Numerically checking hypercontractivity. Certifiable hypercontractivity can be easily assessed numerically using the definition in eq. (28). The definition of k-certifiable hypercontractivity essentially requires that the polynomial

$$h(\boldsymbol{v}) = C(t)^t \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{A}_i^\mathsf{T} \boldsymbol{v} \right\|_2^2 \right)^t - \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{A}_i^\mathsf{T} \boldsymbol{v} \right\|_2^{2t} \right) \tag{41}$$

is a sum-of-squares polynomial for every $t \leq k/2$ and for a given choice of $C(t)^t$. While larger values of k lead to slightly better bounds here we are interested in the smallest choice of k (which is $k = 4$ according to Theorem 12) since the larger the k the larger the size of the resulting moment relaxation. Choosing $k = 4$, we are only left to test that $h(\boldsymbol{v})$ is sos for $t = 2$ (the condition is trivially satisfied for $t = 1$ as long as $C(t)^t \geq 1$). For $t = 2$, we test multiple values of $C(t)^t$ and for each we check if $h(\boldsymbol{v})$ is sos using the `findsos` function in the SOSTOOLS library [120]. The function either returns a decomposition of the polynomial into a sum of squares or reports that the polynomial is not sos.

---

[24]The bound improves for larger β since the set of inliers selected by (TLS) $\mathcal{I}_{\mathsf{TLS}}$ typically shrinks for increasing β and hence it becomes less likely to sample data producing small $\sigma_{\min}(\boldsymbol{A}_{\mathcal{J}})$

(a)

Nr. measurements (n)	$C(t)^t=1$	2	3	4	5	6
5	0	0	18	56	100	100
10	0	0	43	88	99	100
20	0	0	93	99	100	100
40	0	0	99	100	100	100
60	0	3	100	100	100	100
80	0	5	100	100	100	100
100	0	10	100	100	100	100
200	0	70	100	100	100	100

(b)

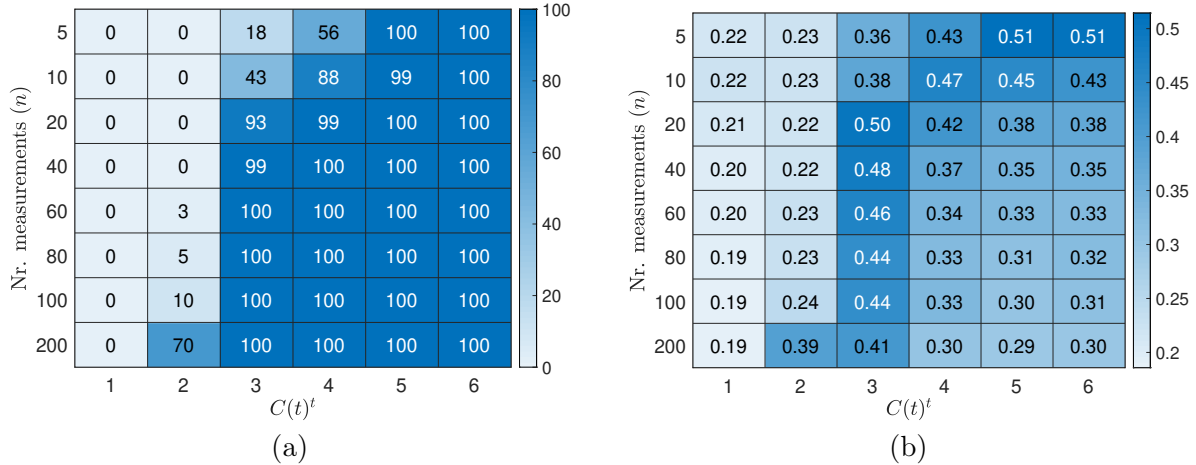| Nr. measurements (n)	$C(t)^t=1$	2	3	4	5	6
5	0.22	0.23	0.36	0.43	0.51	0.51
10	0.22	0.23	0.38	0.47	0.45	0.43
20	0.21	0.22	0.50	0.42	0.38	0.38
40	0.20	0.22	0.48	0.37	0.35	0.35
60	0.20	0.23	0.46	0.34	0.33	0.33
80	0.19	0.23	0.44	0.33	0.31	0.32
100	0.19	0.24	0.44	0.33	0.30	0.31
200	0.19	0.39	0.41	0.30	0.29	0.30

Figure 5: Certifiable hypercontractivity results on synthetic data: (a) Percentage of tests where 4-certifiable hypercontractivity holds for various choices of $C(t)^t$ and number of measurements n. (b) Average runtime (in seconds, averaged over 100 tests) for the same test configurations in figure (a).

We remark that Theorem 12 requires that the set of all n matrices $\boldsymbol{A}_i$ —where potential outliers are replaced with inliers— satisfies k-certifiable hypercontractivity. Therefore, if we assume that the outliers only contaminate the vectors $\boldsymbol{b}_i$ (hence do not alter the $\boldsymbol{A}_i$'s, which are only influenced by the vectors $\boldsymbol{a}_i$'s),[25] we can actually check hypercontractivity in practice. Note that for $t = 2$, $h(\boldsymbol{v})$ is a degree-4 polynomial in the variable $\boldsymbol{v}$ (which is 9-dimensional in our rotation search example): therefore, as we will see, it is quite fast to check if $h(\boldsymbol{v})$ is sos using off-the-shelf solvers.

Experiments on synthetic data. We show that 4-certifiable hypercontractivity is satisfied for small $C(t)^t$ in our synthetic examples and observe that we can numerically check if a given set of measurements is certifiably hypercontractive in a fraction of a second.

Figure 5(a) reports the number of tests (over 100 runs) where 4-certifiable hypercontractivity holds for various choices of $C(t)^t$ (chosen as a constant ranging from 1 to 6) and for increasing number of measurements n (we test $n = \{5, 10, 20, 40, 60, 80, 100, 200\}$). All instances were found to satisfy 4-certifiably hypercontractivity for $C(t)^t = 6$ (independently of the choice of n), while no instance satisfied the property for $C(t)^t = 1$. From Fig. 5(a), we observe a sharp phase transition between the two regimes, and we note that large measurement sets tend to satisfy 4-certifiable hypercontractivity for even smaller values of $C(t)^t$. This is desirable, since the smaller the $C(t)^t$, the better are the error bounds we obtain from Theorem 12.

Figure 5(b) reports the average runtime (in seconds, averaged over 100 tests) for the same test configurations in Fig. 5(a). The average runtime remains below 1 second in all cases and is fairly insensitive to the number of measurements. It is interesting to notice that the solver is faster in resolving instances that are clearly certifiably hypercontractive or not, while it has to work harder to decide the instances at the "boundary" of the phase transition observed in Fig. 5(a).

Experiments on real data. We show that 4-certifiable hypercontractivity also holds for real data for a small value of $C(t)^t$ (with $t = 2$), with a similar trend as the one observed in the synthetic experiments above. Towards this goal, we consider the image stitching data described in Section 8.1. Table 1 reports the percentage of tests where 4-certifiable hypercontractivity holds for

---

[25]This assumption is mild in many rotation search applications, for instance when $\boldsymbol{a}_i$ are read from a database or we have control on how to sample them.

Table 1: Certifiable hypercontractivity results on real image stitching data: percentage of tests where 4-certifiable hypercontractivity holds for increasing $C(2)^2$.

$C(2)^2 =$	1	2	3	4	5	6	7	8	9	10	15	20
% cert. hypercon.	0.0	0.0	0.0	8.3	50.0	58.3	75.0	83.3	83.3	91.7	100.0	100.0

various choices of $C(t)^t$ (chosen as a constant between 1 and 20). Most instances satisfy 4-certifiably hypercontractivity already for $C(t)^t = 6$ and all instances satisfy the property for $C(t)^t = 15$, confirming that certifiable hypercontractivity is a mild assumption for the rotation search problem.

Bound in Theorem 12. We evaluate the bound in Theorem 12 by plotting the coefficients $C_1(k,\beta)^{\frac{2}{k}}$ and $C_2(k,\beta)^{\frac{2}{k}}$ in the theorem, and the maximum outlier rate $\beta_{\max}$ for which the theorem holds (*i.e.*, $\beta_{\max} = \sqrt[\frac{k}{2}-1]{1/(C(k/2)^{\frac{k}{2}} 2^{3k-1})}$). Ideally, we would like for both $C_1(k,\beta)^{\frac{2}{k}}$ and $C_2(k,\beta)^{\frac{2}{k}}$ to be as small as possible, which would make the residual error of our estimate $\boldsymbol{x}_{\mathsf{lts-sdp1}}$ with respect to the inliers $\mathrm{err}_{\dot{\mathcal{I}}}(\boldsymbol{x}_{\mathsf{lts-sdp1}})$ as close as possible to the residual error $\mathrm{opt}_{\dot{\mathcal{I}}}$ an oracle estimator that has access to all the inliers would commit. Moreover, we would like $\beta_{\max}$ to be as large as possible (but still below 0.5), in order for the theorem to be broadly applicable.

Figure 6(a-b) show the coefficients $C_1(k,\beta)^{\frac{2}{k}}$ and $C_2(k,\beta)^{\frac{2}{k}}$ for $k = 4$ and for increasing outlier rates β; we plot multiple lines corresponding to different choices of the coefficient $C(t)^t \in \{1, 2, 4, 6\}$. For each choice of $C(t)^t$ we also plot the maximum outlier rate $\beta_{\max}$ (visualized as a dashed vertical line) for which that error bound holds (indeed it can be seen that the bounds diverge as they approach their admissible upper bound). To interpret the coefficients, the reader might observe that when $C_1(k,\beta)^{\frac{2}{k}} = 1$ it means that the estimator has twice the error of the oracle, while when $C_1(k,\beta)^{\frac{2}{k}} = 10$ it has an error that is more than 10 times larger than the oracle. By observing the x-axis of the plot, we quickly realize that for $k = 4$, even for small values of $C(t)^t$, the maximum outlier rate captured by the theorem is very small (in the order of 10^{-4}); this makes the theorem overly conservative for robotics applications where even in the low-outlier case, the outliers can approach 50%, and where the number of measurements typically ranges from tens to thousands (which, with an outlier rate of 10^{-4}, would only allow for a handful of outliers).

Figure 6(c-d) show the coefficients $C_1(k,\beta)^{\frac{2}{k}}$ and $C_2(k,\beta)^{\frac{2}{k}}$ for $C(t)^t = 6$ and for increasing outlier rates β; in this case, we plot multiple lines corresponding to different choices of $k \in \{4, 6, 8, 10\}$. For each choice of k we also plot the maximum outlier rate $\beta_{\max}$ (visualized as a dashed vertical line). As we mentioned, the bound in the theorem becomes stronger for higher k (at the cost of much increased computation in solving the moment relaxation). At the same time, even with a k as large at 10 the bound remains applicable only to outlier rates below $4.5 \cdot 10^{-3}$, which is unfortunately still tiny to be of practical interest for robotics and vision.

8.4 Low Outlier Rates: A Priori Bounds and Anti-Concentration

This section shows that k-certifiable anti-concentration is a more stringent condition in practice, but it can still be satisfied for $k = 6$ and for suitable choices of η that make the error bounds in Propositions 13 and 14 nontrivial. While in the rotation search problem the property is only satisfied for outlier rates close to zero, we show that on a variant of the problem, certifiable anti-concentration is satisfied for outlier rates up to 50%. To the best of our knowledge, this is the first study evaluating certifiable anti-concentration numerically and showing its applicability with low-degree polynomials p.
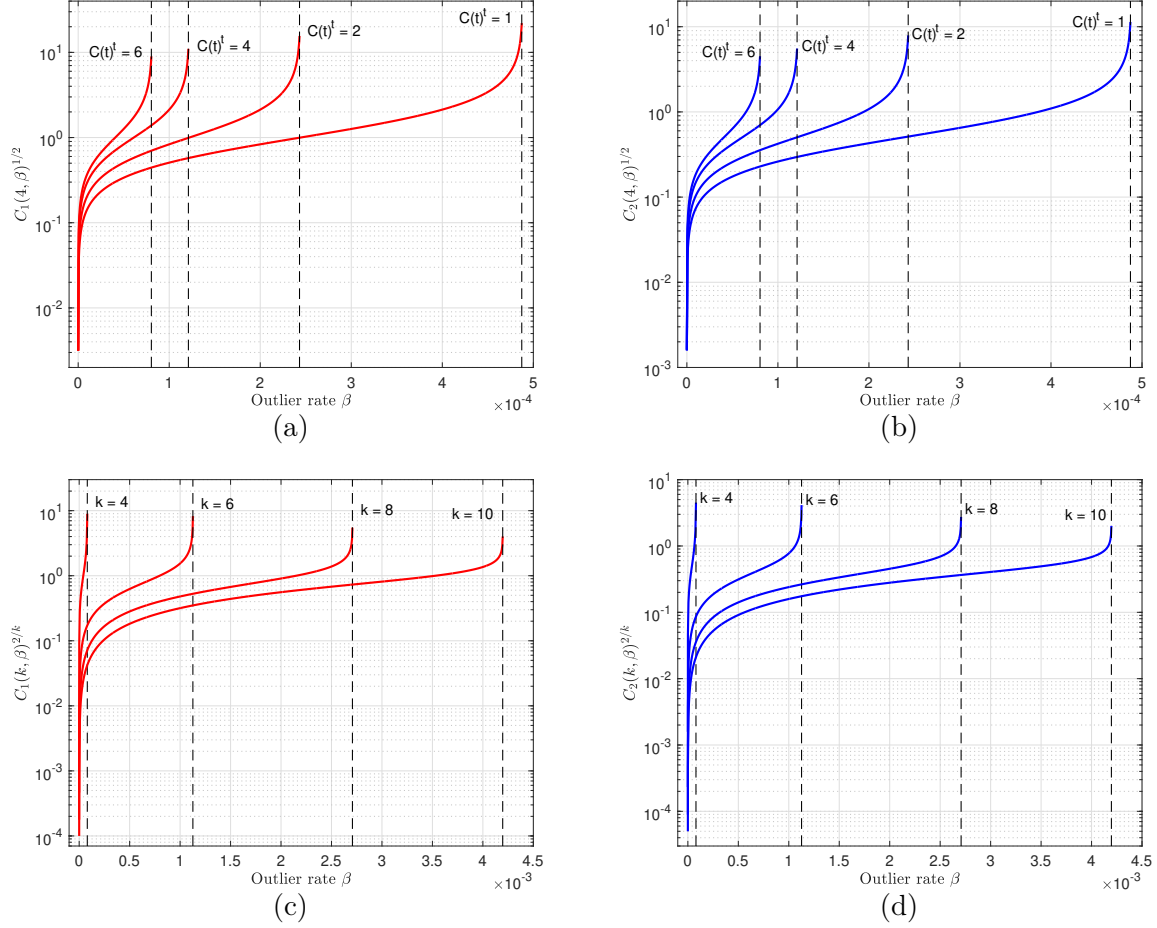
Figure 6: Coefficients $C_1(k,\beta)^{\frac{2}{k}}$ (left column), $C_2(k,\beta)^{\frac{2}{k}}$ (right column), and $\beta_{\max}$ (visualized as dashed vertical lines in each figure) in Theorem 12 for different choices of $C(t)^t$ (sub-figures (a) and (b)) and k (sub-figures (c) and (d)), and for increasing outlier rates.

Numerically checking anti-concentration. Certifiable anti-concentration can be assessed numerically (given the set of inliers $\mathcal{I}$) using the definition in eqs. (32)-(33). Let us rewrite the anti-concentration conditions as two polynomial optimization problems:

$$\min_{\boldsymbol{v}} \quad p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right) - (1-\delta)^2, \quad \text{subject to } \left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2^2 \leq \delta^2 \tag{42}$$

$$\min_{\boldsymbol{v}} \quad C\,\delta\,M^2 - \|\boldsymbol{v}\|^2 \cdot \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right), \quad \text{subject to } \|\boldsymbol{v}\|_2^2 \leq M^2. \tag{43}$$

Anti-concentration requires that the optimal cost of both problems remains greater than zero such that $p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right) \geq (1-\delta)^2$ (for all $i \in \mathcal{I}$) and $\|\boldsymbol{v}\|^2 \cdot \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right) \leq C\,\delta\,M^2$ as requested by the definition. *Certifiable* anti-concentration requires an sos proof of anti-concentration; because of the duality between moment relaxations and sos relaxations, certifiable anti-concentration essentially requires that even if we apply a moment relaxation (of suitable order) to the polynomial optimization problems (42)-(43), then the optimal objective of the relaxation remains positive. Clearly, asking

the moment relaxation to produce a positive objective is a sufficient condition for the objective of (42)-(43) to remain positive, hence certifiable anti-concentration is a more stringent condition.

Given the matrices $\boldsymbol{A}_i$ (for all the inliers), the polynomial p, and the coefficients C, δ, M, and η, we can numerically solve a moment relaxation of (42)-(43) using the `findbound` function in the SOSTOOLS library [120]. We note that (42) involves polynomials of degree up to $2 \cdot \deg(p)$ while (43) involves polynomials of degree up to $2 \cdot \deg(p) + 2$, both in the variable $\boldsymbol{v}$ (which is 9-dimensional in our rotation search example). In practice, the `findbound` function runs out of memory for degrees larger than 8 in our problem, which imposes to choose a degree-2 polynomial p. For such choice of p, (42) involves degree-4 polynomials, while (43) involves degree-6 polynomials, which has implications on the minimum order of the moment relaxation we compute via `findbound`.[26]

So far we concluded that we can numerically check certifiable anti-concentration given the matrices $\boldsymbol{A}_i$, the polynomial p, and the coefficients C, δ, M, and η. The matrices $\boldsymbol{A}_i$ are given as an input to the problem, while our theorems (Propositions 13 to 15) specify how to set δ, C, and M. Therefore, we are only left to figure our how to set the constant η and the polynomial p. As we will see, these parameters control the trade-off between tightness of our error bound and computational cost (more precisely, the order of the moment relaxation) of the corresponding estimators.

Setting p and η. The goal of this subsection is to provide insights on the selection of p and η. As we have already observed, for computational reasons we are going to restrict ourselves to degree-2 polynomials p; however later in this subsection we are also going to comment on the impact of higher-degree polynomials p. Definition 11 imposes $p(0) = 1$ and requires p to be an even polynomial, hence a degree-2 polynomial p is restricted to the form:

$$p(a) = 1 + c_2 a^2, \tag{44}$$

for some scalar variable a and coefficient c_2. In order to evaluate potential choices of c_2 we consider the case of a single inlier ($|\mathcal{I}| = 1$) and visualize the functions involved in the definition of certifiable anti-concentration. In particular, we observe that a necessary condition for certifiable anti-concentration to be satisfied is that the following inequalities hold in the traditional sense:[27]

$$p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right) \geq (1 - \delta)^2, \quad \forall \boldsymbol{v} \text{ such that } \left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2^2 \leq \delta^2 \tag{45}$$

$$p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right) \leq \frac{C\,\delta\,M^2}{\|\boldsymbol{v}\|^2}, \quad \forall \boldsymbol{v} \text{ such that } \|\boldsymbol{v}\|_2^2 \leq M^2, \tag{46}$$

where we assumed $|\mathcal{I}| = 1$. Now we see that the conditions (45)-(46) can be understood as constraints on the behavior of the polynomial $p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right)$. Figure 7(a) shows the behavior of the functions involved in the inequalities (45)-(46). We visualize 4 choices of $p^2\left(\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2\right)$, corresponding to 4 different choices of coefficients $c_2 \in \{-1, -0.3, -0.1, -0.05, -0.01\}$ in (44). We also visualize the bounds $(1 - \delta)^2$ in (45), which only has to hold for $\left\|\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{v}\right\|_2^2 \leq \delta^2$. Moreover, we visualize the upper bound $\frac{C\,\delta\,M^2}{\|\boldsymbol{v}\|^2}$ for C, δ, M chosen as in Proposition 13 and for $\eta = 3.26$ (we discuss how to choose η below). Note that the latter function has a different x-axis ($\|\boldsymbol{v}\|$, in red), but, noting that the matrices $\boldsymbol{A}_i$ in the rotation search problem have maximum singular value equal to 1, the

---

[26]`findbound` takes a "`degree`" input argument, that can be understood as the level of the pseudo-distribution computed by the function. In this case, we use the minimum `degree`=4 for (42) and `degree`=6 for (43).

[27]Recall again that asking for an sos proof imposes a stronger condition compared to just asking for the inequality to be satisfied in the traditional sense.
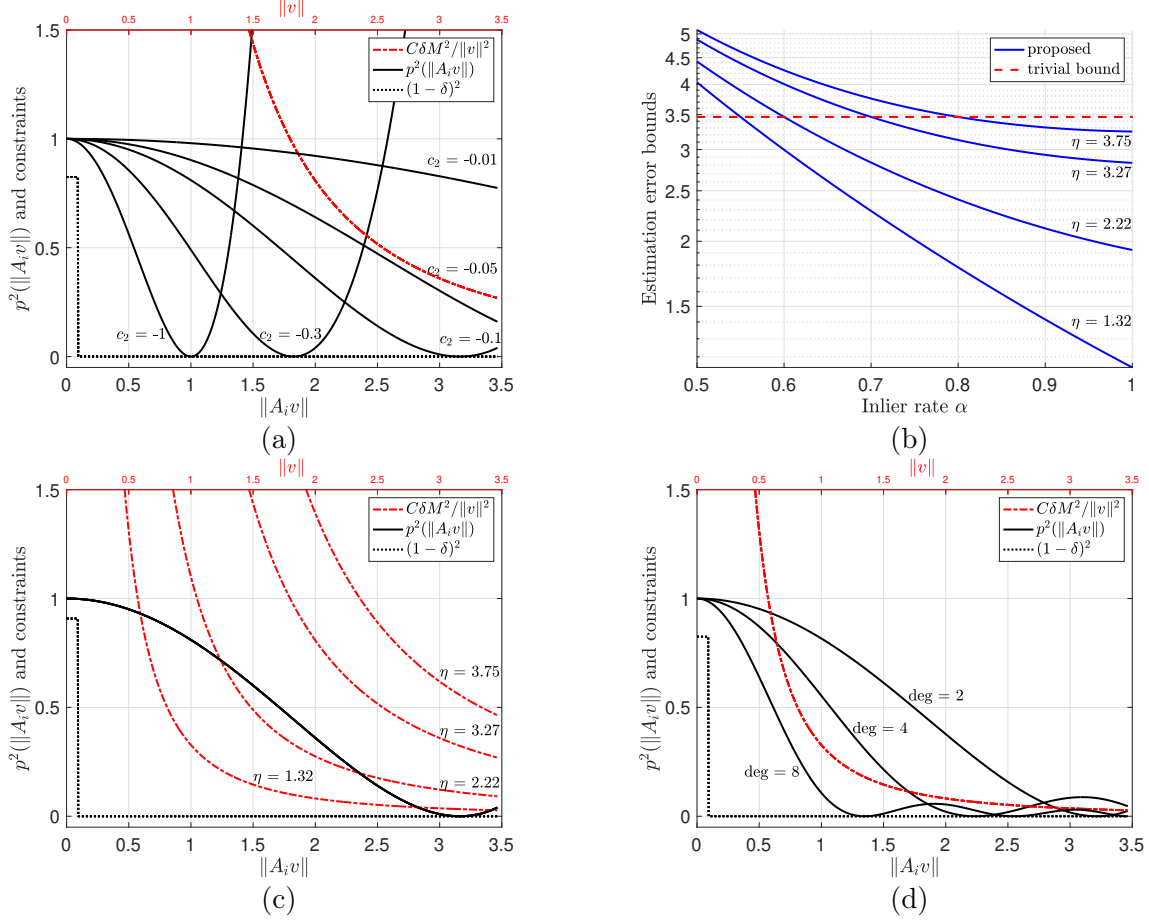
Figure 7: Visualization of the functions involved in the definition of anti-concentration in eq. (45)-(46). (a) Polynomials p^2 parametrized as in (44) for different choices of coefficient c_2 compared to the bounds in (45)-(46) for $\eta = 3.26$. (b) Estimation error bounds from Propositions 13 and 14 for different choices of η. (c) Polynomials p^2 parametrized as in (44) with $c_2 = -0.1$, and bounds in (45)-(46) for different choices of η. (d) Polynomials p^2 of degree 2, 4, 8, and suitable choices of coefficients compared to the bounds in (45)-(46) for $\eta = 1.32$.

axes are still useful to visualize together.[28] In this example with $|\mathcal{I}| = 1$, a necessary condition for certifiable anti-concentration to hold is that the squared polynomial p^2 remains below the red upper bound in Fig. 7(a) and above the lower bounds $(1 - \delta)^2$ visualized as a black dotted line in the figure. From the figure it is clear that the latter condition is relatively easy to satisfy (at least for degree-2 polynomials), while remaining below the red upper bound is a more stringent constraint. In particular, if we choose large negative values of the coefficient c_2 (e.g., $c_2 = -1$ or $c_2 = -0.3$), the squared polynomial p^2 "bounces back" and quickly reaches the upper bound. If the coefficient is too small (e.g., $c_2 = -0.01$ in the figure), the polynomial does not decrease sufficiently fast and still

---

[28]Since the matrices $\boldsymbol{A}_i$ are built from unit vectors, it can be easily shown that the matrices have the largest singular value equal to 1. Therefore, for each choice of $\|\boldsymbol{v}\|$ one can choose the direction of $\boldsymbol{v}$ to make $\|\boldsymbol{A}_i\boldsymbol{v}\| = \|\boldsymbol{v}\|$, making the top and bottom x-axes of Fig. 7(a) commensurable; however note that $\|\boldsymbol{A}_i\boldsymbol{v}\| \leq \|\boldsymbol{v}\|$ in general, and $\|\boldsymbol{A}_i\boldsymbol{v}\|$ can become as small as zero for rank-deficient $\boldsymbol{A}_i$.

violates the upper bound. However, choosing $c_2 = -0.1$ and $c_2 = -0.05$ satisfies both constraints, with $c_2 = -0.1$ being the farthest of the two from the upper bound. Later on, we will show that $c_2 = -0.1$ indeed leads to certifying anti-concentration in our rotation search problem for a suitably chosen η. We remark that while Fig. 7(a) is somewhat insightful in visualizing the constraints that the polynomial p has to satisfy, when $|\mathcal{I}| > 1$ the upper bound has to be only satisfied in average (*i.e.*, $\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p^2 \left(\left\| \boldsymbol{A}_i^\mathsf{T} \boldsymbol{v} \right\|_2 \right) \leq \frac{C \delta M^2}{\|\boldsymbol{v}\|^2}$ rather than $p^2 \left(\left\| \boldsymbol{A}_i^\mathsf{T} \boldsymbol{v} \right\|_2 \right) \leq \frac{C \delta M^2}{\|\boldsymbol{v}\|^2}$ for all i), giving some extra slack in satisfying anti-concentration.

Now let us turn our attention to the parameter η. The parameter η can be roughly understood as a measure of the "quality" of the inlier set: good inlier sets will be certifiably anti-concentrated for a small η, which induces better error bounds (all bounds in Propositions 13 to 15 increase monotonically with η). On the other hand, while we can arbitrarily increase the η to make a given batch of inliers certifiably anti-concentrated, a large η might lead to a trivial bound in the corresponding theorem. In particular, for the bounds in Propositions 13 and 14 to be nontrivial (*i.e.*, smaller than the trivial bound $2M_x$), we need:

$$\left(\frac{\alpha \, \eta}{2} + 2 \frac{1 - \alpha}{\alpha} \right) < 2 \quad \implies \quad \eta < \frac{2}{\alpha} \left(2 - 2 \frac{1 - \alpha}{\alpha} \right). \tag{47}$$

Therefore, for a certain inlier rate α, we can choose the η such that our theorems do not produce a trivial bound and then test is such η ensures certifiable anti-concentration. Ideally, we would like to have a nontrivial bound already around $\alpha = 0.5$, but as we will see in a second that might lead to a small η, which makes the certifiable anti-concentration condition too strict. Below, we choose potential values of η according to (47), such that they produce nontrivial bounds for $\alpha \in [0.55, 0.6, 0.7, 0.8]$, respectively leading to the $\eta \in [1.32, 2.22, 3.26, 3.75]$. Figure 7(b) shows the trend of the bounds in Propositions 13 and 14 for those choices of η, compared with the trivial bound $2M_x = 2\sqrt{3}$. As expected, our choices of η are designed such that the proposed bounds move below the trivial bound for $\alpha \in [0.55, 0.6, 0.7, 0.8]$, respectively. Ideally, we would like to choose the smallest possible η. Figure 7(c) visualizes the price we need to pay to decrease the η: the figure shows how smaller values of η lead to increasingly tight upper bounds $\frac{C \delta M^2}{\|\boldsymbol{v}\|^2}$ that the polynomial p needs to satisfy. The figure already shows that in this case ($|\mathcal{I}| = 1$) it is unlikely we can find a degree-2 polynomial that satisfies the upper bound for $\eta = 1.32$. Figure 7(d) shows that we can still satisfy the upper bound $\frac{C \delta M^2}{\|\boldsymbol{v}\|^2}$ for a small η if we are willing to increase the degree of the polynomial p: the figure plots even polynomials of degree 2, 4, and 8 (with suitably chosen coefficients) contrasting them with the upper bound obtained for $\eta = 1.32$. It can be observed that the degree-8 polynomial remains below the upper bound in the figure. This shows that we can afford stronger guarantees (*i.e.*, smaller η) if we are willing to pay the computational cost of using a higher-degree polynomial, which in turns implies a larger k in the definition of certifiable anti-concentration and a higher order of the moment relaxations involved in our robust estimators. In the following we stick to the more realistic case of degree-2 polynomial p and we numerically investigate which η we can "afford" in synthetic problems for both the rotation search problem and for a variation of the problem.

Experiments on synthetic data. We choose the univariate polynomial to be $p(a) = 1 - 0.1a^2$ and test certifiable anti-concentration in a rotation search problem with $n = 50$ measurements, generated as described in Section 8.1. Testing certifiable anti-concentration is much slower than testing certifiable hypercontractivity, hence we only compute statistics over 10 runs. We test whether certifiable anti-concentration holds for increasing outlier rates $\beta \in \{0; 0.1; 0.3; 0.5; 0.7; 0.9\}$ and for different choices of $\eta \in \{1.32; 2.22; 3.26; 3.75\}$. Fig. 8(a) reports the percentage of tests where

42

Figure 8(a) — η vs Outlier rate (β), n = 50

η	0	0.1	0.3	0.5	0.7	0.9
1.32	0.00	0.00	0.00	0.00	0.00	0.00
2.22	0.00	0.00	0.00	0.00	0.00	0.00
3.27	20.00	0.00	0.00	0.00	0.00	0.00
3.75	100.00	70.00	0.00	0.00	0.00	0.00

Figure 8(b) — Nr. measurements (n) vs Outlier rate (β), η = 3.75

n	0	0.1	0.3	0.5	0.7	0.9
10	70.00	0.00	0.00	0.00	0.00	0.00
50	100.00	80.00	0.00	0.00	0.00	0.00
100	100.00	80.00	0.00	0.00	0.00	0.00
200	100.00	100.00	0.00	0.00	0.00	0.00

Figure 8(c) — η vs Outlier rate (β), n = 50

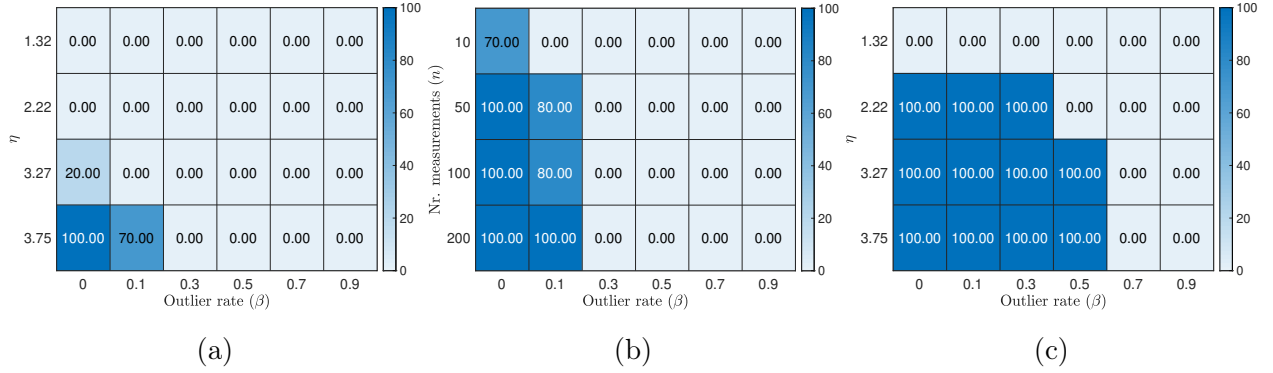| η	0	0.1	0.3	0.5	0.7	0.9
1.32	0.00	0.00	0.00	0.00	0.00	0.00
2.22	100.00	100.00	100.00	0.00	0.00	0.00
3.27	100.00	100.00	100.00	100.00	0.00	0.00
3.75	100.00	100.00	100.00	100.00	0.00	0.00

(a)   (b)   (c)

Figure 8: Certifiable anti-concentration results on synthetic data: (a) Rotation search problem: percentage of tests where 6-certifiable anti-concentration holds for various choices of η and outlier rates β (with $n = 50$). (b) Rotation search problem: percentage of tests where 6-certifiable anti-concentration holds for increasing number of measurements n and outlier rates β (with $\eta = 3.75$). (b) Modified version of the rotation search problem: percentage of tests where 6-certifiable anti-concentration holds for various choices of η and β (with $n = 50$). Statistics computed over 10 runs.

6-certifiable anti-concentration holds for various choices of η and outlier rates β. We observe that certifiable anti-concentration is only satisfied for a relatively large η and for small outlier rates. In particular, the parameter C in (46) depends on the square of $\alpha = 1 - \beta$ according to Propositions 13 and 14, hence quickly making the condition too strict for increasing outlier rates β.

Figure 8(b) reports the percentage of tests where 6-certifiable anti-concentration holds for increasing number of measurements n and outlier rates β, and for $\eta = 3.75$. The figure shows that certifiable anti-concentration is more likely to hold for larger sets of measurements, while still only applying to problems with relatively low outlier rates (*i.e.*, 10%). While Figure 8(a-b) seem to suggest that certifiable anti-concentration with degree-2 polynomials p is too strict of a condition to be of practical interest, below we report a variant of the rotation search problem, where the condition is satisfied for much larger outlier rates β.

We consider a modified version of the rotation search problem, where each measurement, instead of being a single 3D vector $\boldsymbol{b}_i$, is a triplet of orthogonal vectors $\boldsymbol{b}_i^{(1)}$, $\boldsymbol{b}_i^{(2)}$, $\boldsymbol{b}_i^{(3)}$. In this variant of the rotation search problem, the measurement model becomes:

$$\boldsymbol{y}_i = \begin{bmatrix} \boldsymbol{b}_i^{(1)} \\ \boldsymbol{b}_i^{(2)} \\ \boldsymbol{b}_i^{(3)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}\boldsymbol{a}_i^{(1)} + \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{R}\boldsymbol{a}_i^{(2)} + \boldsymbol{\epsilon}^{(2)} \\ \boldsymbol{R}\boldsymbol{a}_i^{(3)} + \boldsymbol{\epsilon}^{(3)} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{a}_i^{(1)})^\mathsf{T} \otimes \mathbf{I}_3 \\ (\boldsymbol{a}_i^{(2)})^\mathsf{T} \otimes \mathbf{I}_3 \\ (\boldsymbol{a}_i^{(3)})^\mathsf{T} \otimes \mathbf{I}_3 \end{bmatrix} \cdot vec(\boldsymbol{R}) = \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ + \boldsymbol{\epsilon}. \qquad (48)$$

The expert reader might notice that this is now similar to a *rotation averaging* problem (see Example 18 in Section 9), since each measurement allows reconstructing the full rotation $\boldsymbol{R}$. We numerically test certifiable anti-concentration in this problem variant. Towards this goal, we use the same data generation protocol described in Section 8.1 to sample $\boldsymbol{a}_i$, and set $\boldsymbol{a}_i^{(1)} = \boldsymbol{a}_i$; then we create an orthogonal vector $\boldsymbol{a}_i^{(2)}$ by sampling a second random vector, projecting it onto the null space of $\boldsymbol{a}_i^{(1)}$, and normalizing it to have unit norm; finally, we set $\boldsymbol{a}_i^{(3)} = \boldsymbol{a}_i^{(1)} \times \boldsymbol{a}_i^{(2)}$, where $\times$ is the cross product. Note that the vectors $\boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}, \boldsymbol{b}_i^{(3)}$ do not appear in the definition of anti-concentration,

hence we do not have to generate them. Figure 8(c) reports the percentage of tests where 6-certifiable anti-concentration holds for various choices of η and outlier rates β in this variant of the rotation search problem. In this case, certifiable anti-concentration is more broadly satisfied for several choices of η and for outlier rates as high as 50%. To the best of our knowledge, this is the first numerical evidence supporting the use of certifiable anti-concentration with degree-2 polynomials p. Moreover, since certifiable anti-concentration holds for relatively small η, this variant of the problem enjoys the tighter error bounds shown in Fig. 7(b). Besides the intuition that this variant of the rotation search problem is "easier" (an aspect that is correctly captured by η for which certifiable anti-concentration holds), there is a precise technical reason that makes the property easier to satisfy. By inspection, we note that the matrix $\boldsymbol{A}_i$ in eq. (48) is an orthogonal matrix,[29] hence $\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\| = \|\boldsymbol{v}\|$. This makes the top and bottom x-axes in Fig. 7(a) the same, and the anti-concentration condition becomes a point-wise comparison between the polynomial p^2 and the bounds.[30]

The runtime to check certifiable anti-concentration is much higher than the one required for certifiable hypercontractivity, since now we have to solve constrained optimization problems, also involving higher-degree polynomials. Solving a moment relaxation of (42) using $\texttt{findbound}$ [120] takes around 0.5 seconds on a standard laptop. However, solving a relaxation of (43) requires around 5 minutes, due to the need to use a higher-order relaxation (recall that (43) involves degree-6 polynomials).

8.5 High Outlier Rates: List-Decodable Estimation

We conclude this experimental section by discussing list-decodable estimation and the high-outlier regime. In particular, this section shows that despite the fact that the theoretical guarantees of Theorem 17 fall short in practice, a small variation of Algorithm 5 has impressive empirical performance. To the best of our knowledge, the algorithm we describe is the first practical algorithm for list-decodable estimation based on the framework proposed in [37].

We start by observing that the theoretical results in Theorem 17 fall short of providing usable performance guarantees in the high-outlier-rate case: first of all, certifiable anti-concentration, even in the simpler variant of the rotation search problem in Fig. 8(c), does not empirically hold for outlier rates $\beta > 0.5$ (at least when using a degree-2 p), hence the results only cover the low-outlier regime in practice. Second, the guarantees in Theorem 17 (*i.e.*, the bound $\|\boldsymbol{x} - \boldsymbol{x}^\circ\|_2 \leq \eta M_x$) are non-trivial only for $\eta < 2$; however, Fig. 8(c) shows that anti-concentration starts holding for $\eta > 2$ in most cases, making the result in Theorem 17 of pure theoretical interest. Finally, the smallest k allowed by the theorem is $k = 6$, which leads to large order-3 moment relaxations that we cannot solve with current SDP solvers. Despite these limitations, one might still wonder if Algorithm 5 would produce good estimates even for lower relaxation orders, beyond what's covered by the theory.

This section shows that a sparse order-2 relaxation of problem (LDR) in Algorithm 5 already leads to accurate list-decodable estimation. In particular, we show that a small variant of Algorithm 5, that we call *Sparse LIst-Decodable Estimation* (SLIDE), (i) produces good estimates for problems

---

[29] $\boldsymbol{A}_i^\mathsf{T}\boldsymbol{A}_i = \begin{bmatrix} (\boldsymbol{a}_i^{(1)})^\mathsf{T} \otimes \mathbf{I}_3 \\ (\boldsymbol{a}_i^{(2)})^\mathsf{T} \otimes \mathbf{I}_3 \\ (\boldsymbol{a}_i^{(3)})^\mathsf{T} \otimes \mathbf{I}_3 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{a}_i^{(1)} \otimes \mathbf{I}_3 \ \boldsymbol{a}_i^{(2)} \otimes \mathbf{I}_3 \ \boldsymbol{a}_i^{(3)} \otimes \mathbf{I}_3 \end{bmatrix} = \begin{bmatrix} (\boldsymbol{a}_i^{(1)})^\mathsf{T}\boldsymbol{a}_i^{(1)}\mathbf{I}_3 & (\boldsymbol{a}_i^{(1)})^\mathsf{T}\boldsymbol{a}_i^{(2)}\mathbf{I}_3 & (\boldsymbol{a}_i^{(1)})^\mathsf{T}\boldsymbol{a}_i^{(3)}\mathbf{I}_3 \\ \star & (\boldsymbol{a}_i^{(2)})^\mathsf{T}\boldsymbol{a}_i^{(2)}\mathbf{I}_3 & (\boldsymbol{a}_i^{(2)})^\mathsf{T}\boldsymbol{a}_i^{(3)}\mathbf{I}_3 \\ \star & \star & (\boldsymbol{a}_i^{(3)})^\mathsf{T}\boldsymbol{a}_i^{(3)}\mathbf{I}_3 \end{bmatrix} = \mathbf{I}_9$

[30] As we observed, in the standard rotation search problem, $\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\|$ and $\|\boldsymbol{v}\|$ are different in general, but it still holds $\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\| \leq \|\boldsymbol{v}\|$ since the largest singular value of $\boldsymbol{A}_i$ is equal to 1. Since, however, for certain choices of $\boldsymbol{v}$, $\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\|$ can be much smaller than $\|\boldsymbol{v}\|$ (and be even zero if the matrix $\boldsymbol{A}_i$ is rank-deficient), the condition (46) may need to be satisfied for any $0 \leq \|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\| \leq \|\boldsymbol{v}\|$. On the other hand, when the matrix $\boldsymbol{A}_i$ is orthogonal, $\|\boldsymbol{A}_i^\mathsf{T}\boldsymbol{v}\| = \|\boldsymbol{v}\|$ and the condition in (46) simplifies to $p^2 (\|\boldsymbol{v}\|_2) \leq \frac{C\,\delta\,M^2}{\|\boldsymbol{v}\|^2}$, which is less stringent to satisfy.

with high rates of random outliers, (ii) produces good estimates for problems with high rates of "adversarial" (more precisely, mutually consistent) outliers, and (iii) if the measurements are generated by multiple rotations (*i.e.,* different subsets of measurements are generated by different rotations), then SLIDE simultaneously recovers *all* the rotations generating the data.

Sparse LIst-Decodable Estimation (SLIDE). The proposed algorithm entails three main changes to Algorithm 5, for $r = 2$. First of all, instead of solving a moment relaxation of order $r = 2$, which is still expensive for large n (we test for $n = 50$), we develop a sparse relaxation, whose details are given in Appendix J. The relaxation essentially considers a sparse monomial basis, which neglects certain degree-2 monomials (*e.g.,* $\omega_i \cdot \omega_j$) that would make the dimension of the resulting SDP much larger; the monomial basis is designed to still give us access to the entries of the pseudo-moment matrix used in Algorithm 5. The second modification is to round the list of solutions of Algorithm 5 to the domain $\mathbb{X}$. The latter is a consequential change: we empirically noticed that the original approach in Algorithm 5 (but with our sparse relaxation) produces estimates with norm close to zero. As we will see, projecting the estimates to $\mathbb{X}$ has the effect of re-normalizing the result and correcting scaling problems. Finally, we always return n hypotheses, rather than sampling; this makes the results deterministic and independent from an arbitrary choice of number of hypotheses (which can no longer be guided by the guarantees in Theorem 17). Below, we will refer to this modified algorithm as "SLIDE" (Sparse LIst-Decodable Estimation).

Figure 9: Rotation search with random outliers: (a) Rotation error in degrees for the TLS approach from [34] and SLIDE for increasing outlier rates β. (b) Relaxation gap for the same approaches (y-axis is on log scale). Shaded areas correspond to the 25th, 50th, 75th, and 90th percentiles. Results are averaged over 10 runs.

Experiments on synthetic data. Figure 9 reports statistics comparing the estimation errors of the sparse relaxation of the (TLS) problem proposed in [34] (label: "TLS") against SLIDE, using the data generation protocol of Section 8.1 with $n = 50$ and increasing outlier rates β between 0.1 and 0.9 at increments of 0.1. Figure 9(a) reports the rotation error in degrees between the estimate from each approach and the ground truth. For SLIDE, we report the smallest error across the estimates in the list. The TLS results confirm the observation from [34] that TLS is extremely robust to extreme rates of non-adversarial outliers (as described in Section 8.1, in this case we sample the outliers at random). However, the interesting observation is that SLIDE is also able to retrieve good

estimates in practice as long as the amount of inliers is sufficiently large; note that for $\beta = 0.9$ we only have 5 inliers, which is comparable to the size of the minimum set of measurements we need to uniquely solve the problem ($\bar{d} = 3$ in rotation search). Figure 9(b) reports the relative relaxation gap for both approaches, computed as in eq. (24) in [30]:

$$\gamma = \frac{|f_{\mathrm{sdp}} - \hat{f}|}{1 + |f_{\mathrm{sdp}}| + |\hat{f}|} \tag{49}$$

where f_{sdp} is the optimal objective of the SDP relaxation and $\hat{f}$ is the objective attained by a rounded solution.[31] As already reported in [34], the TLS relaxation remains tight across the spectrum. On the other hand, SLIDE's relaxation is typically loose, while still returning accurate solutions. The lack of tightness of SLIDE's relaxation is a feature rather than a bug: a tight relaxation would produce a binary vector $\boldsymbol{\omega}$,[32] forcing the objective to be always equal to $\|\boldsymbol{\omega}\|_2^2 = \sum_{i=1}^n \omega_i^2 = \sum_{i=1}^n \omega_i = \alpha\,n$; such a relaxation would always produce a set of $\alpha\,n$ inliers rather than "spreading" across multiple potential sets of inliers (*cf.* Section 7). Indeed, in our example in eq. (39), enabling non-binary $\boldsymbol{\omega}$ was the key to allow solutions that spread across multiple potential sets of inliers.
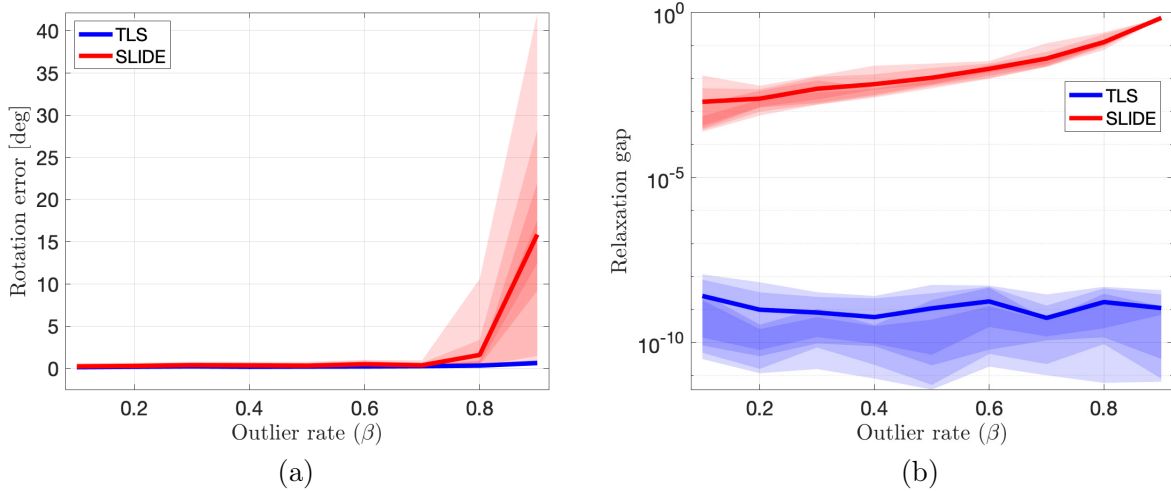

Figure 10: Rotation search with mutually consistent outliers: (a) Rotation error in degrees for the TLS approach from [34] and SLIDE for increasing outlier rates β. (b) Relaxation gap for the same approaches (y-axis is on log scale). Shaded areas correspond to the 25th, 50th, 75th, and 90th percentiles. Results are averaged over 10 runs.

Figure 10 reports the same statistics, but now for a case with mutually consistent outliers. In this case, we generate the set of outliers according to the same generative model of the rotation search model (4), but using a rotation $\boldsymbol{R}' \neq \boldsymbol{R}^\circ$ that we choose at random. Essentially, the set of measurements now contains two different hypotheses for the parameter we want to estimate, and if the outliers are more than 50% any single-hypothesis estimator is expected to fail (and return an estimate close to $\boldsymbol{R}'$ instead of $\boldsymbol{R}^\circ$). Figure 10(a) reports the rotation error for both TLS and SLIDE. As expected, TLS experiences a phase transition, where for $\beta < 0.5$ it is able to retrieve a good

---

[31]For SLIDE, the relaxation gap can be easily computed by observing that the optimal objective of the non-relaxed polynomial optimization problem (LDR) is always $\alpha \cdot n$.

[32]More precisely, it would produce binary entries $\boldsymbol{X}_{[\boldsymbol{\omega}]} \in \{0; 1\}^n$ in the pseudo-moment matrix.

estimate for $\boldsymbol{R}^\circ$, while for $\beta > 0.5$, the outliers form the most likely hypothesis, leading the estimator to perform arbitrarily poorly. However, one can observe how SLIDE still performs similarly to the non-adversarial case, achieving much smaller errors (the mean error remains below $4°$ for $\beta = 0.8$). Figure 9(b) reports the relaxation gap for both approaches. SLIDE's relaxation remains loose, while, quite interestingly, TLS remains tight even in the presence of mutually consistent outliers.

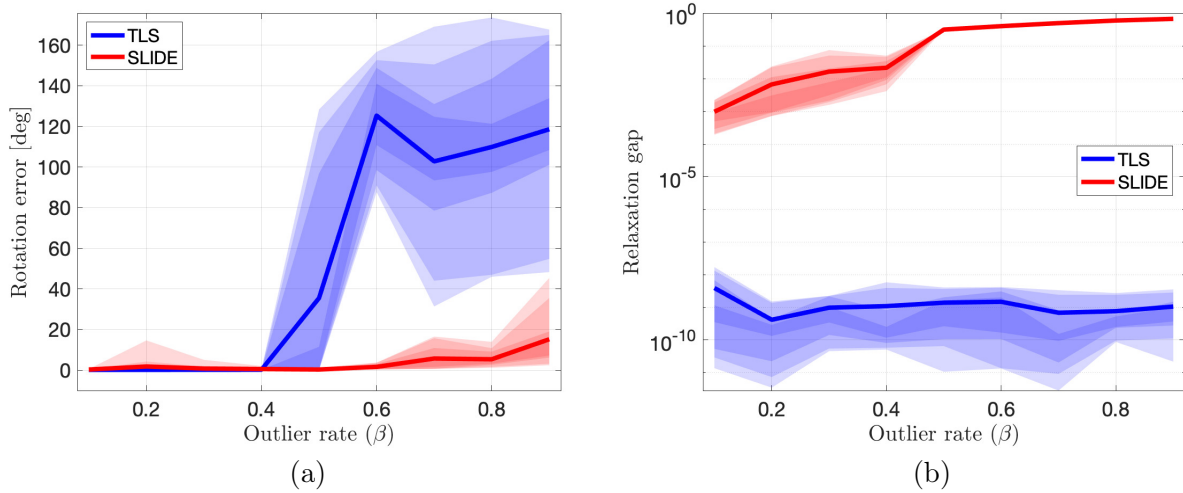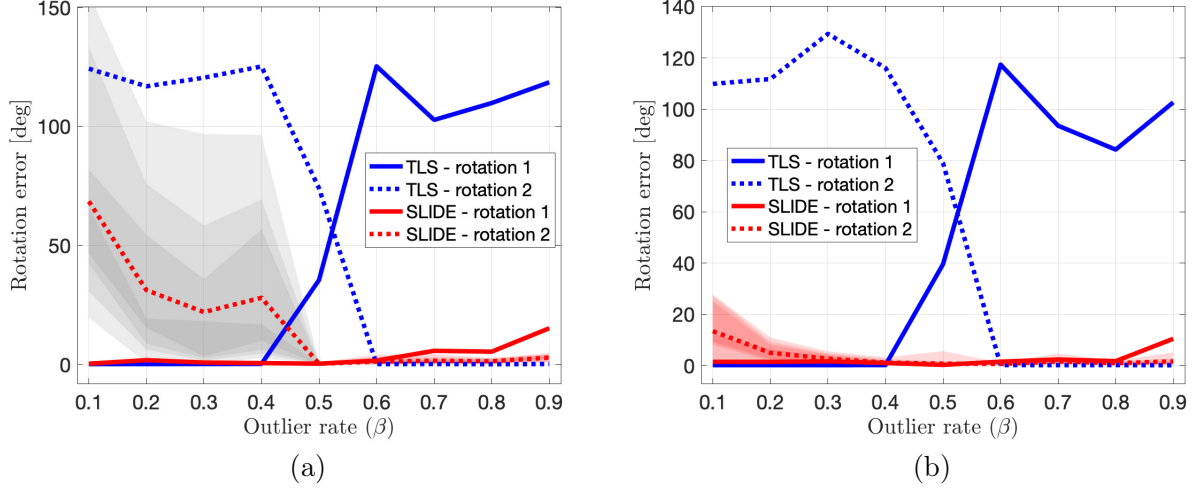
Figure 11: Rotation search with mutually consistent outliers: rotation errors in degrees for the TLS approach from [34] and SLIDE for increasing outlier rates β. Errors are reported for both rotation hypotheses present in the data. Shaded areas are only shown for SLIDE and for the second rotation (the shaded area for the first rotation can be observed in Fig. 9(a)) and correspond to the 25th, 50th, 75th, and 90th percentiles. Results are averaged over 10 runs.

For the same set of experiments with mutually consistent outliers, Fig. 11(a) provides error statistics for both rotations $(\boldsymbol{R}', \boldsymbol{R}^\circ)$ present in the sensor data. As expected, TLS, depending on the outlier rate, retrieves either one or the other rotation (*i.e.,* it only retrieves the dominant hypothesis in the data). Indeed, (TLS) —similarly to (MC)— looks for the estimate supported by the "largest" set of measurements, hence it is not able to find more complex patterns in the measurements. On the other hand, SLIDE's behavior is much more intriguing. Let us start by considering the high-outlier regime ($\beta > 0.5$), which is easier to parse. It is clear that when $\beta > 0.5$, SLIDE is able to report both rotations producing the measurements, as long as the number of inliers for each is sufficiently large (similarly to what we observed in Figs. 9 and 10). This is quite relevant for practical applications: for instance, in rotation search, the presence of different rotations generating the data may be due to the 3D vectors experiencing different motions over time (*e.g.,* a subset belongs to a static portion of the environment, while another subset belongs to a moving object), hence being able to retrieve both might be useful to simultaneously infer motion of multiple objects with respect to the sensor. In the low-outlier case $\beta < 0.5$, it might seem that SLIDE is not able to recover the second rotation; however, this is only a byproduct of our experimental setup: for each β, the polynomial optimization (LDR) that SLIDE solves only searchers for estimates "supported" by at least $\alpha = 1 - \beta$ fraction of the measurements. Therefore, for $\beta = 0.1$, we are asking SLIDE to report any estimate that is supported by at least 90% of the measurements, hence SLIDE correctly reports a single estimate ($\boldsymbol{R}^\circ$), since the other estimate ($\boldsymbol{R}'$) is only supported by 10% of the measurements (*i.e.,* the

outliers). If we instead pass the minimum $\min(\beta, 1 - \beta)$ between the inlier and the outlier rate as the fraction of the measurement set SLIDE has to look for, we obtain the behavior in Fig. 11(b), where SLIDE is able to recover both rotations as long as they are supported by enough measurements.

Figure 12 provides qualitative evidence to further support the observation that SLIDE is empirically able to report all the estimates supported by at least $\alpha\, n$ inliers, as long as $\alpha\, n$ is sufficiently large. Figure 12(a) shows an example with mutually consistent outliers and $\beta = 0.5$: the figure shows the rotations $\boldsymbol{R}^\circ$ and $\boldsymbol{R}'$ generating the data as coordinate frames with thick and short axes, and the estimates in the list returned by SLIDE as coordinate frames with thin and long axes. The figure shows that all the estimates returned by SLIDE cluster around the ground-truth rotations generating the data. Figure 12(b) shows a more extreme case where the data is generated by 5 different rotations, each one producing 20% of the measurements. The figure shows the 5 ground-truth rotations with thick and short axes, and the estimates from SLIDE as thin and long axes. Also in this case, SLIDE's estimates cluster around the ground-truth rotations, and indeed SLIDE is able to recover all rotations within a maximum error of $2.6°$. These results also seem to suggest other algorithmic variants of Algorithm 5, where one, rather than sampling a smaller list, would get a list of size n and then reduce the size of the list by clustering it into fewer hypotheses. We discuss this and other extensions in the following section.

As we mentioned in Section 8.1, our implementation of SLIDE is in Matlab and uses MOSEK [105] as an SDP solver. The average runtime of SLIDE in problems with $n = 50$ is 3 minutes.

(a)   (b)
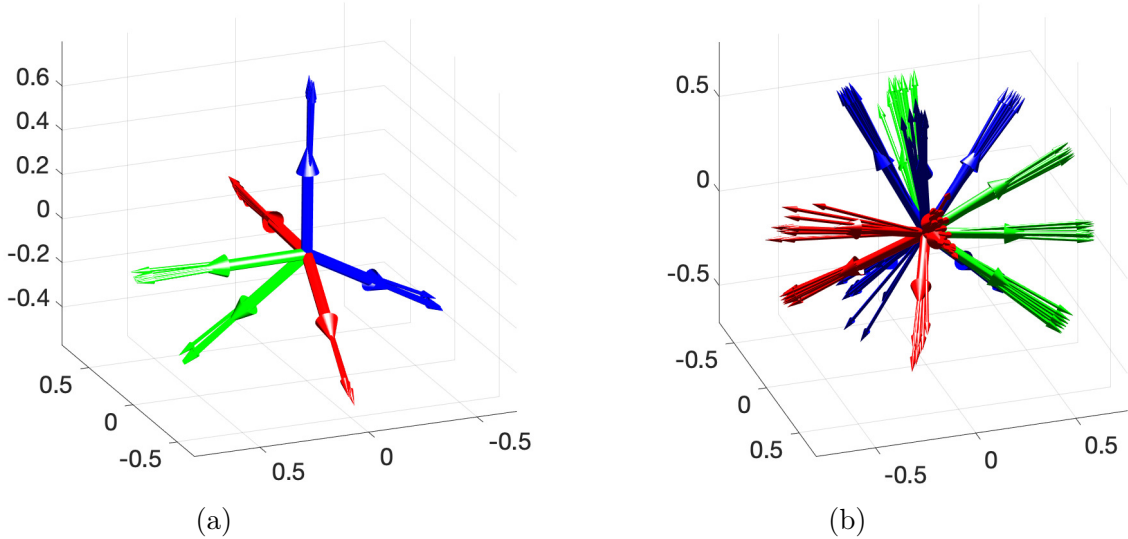
Figure 12: Ground-truth rotations generating the data vs. estimates from SLIDE. The ground-truth rotations are visualized as coordinate frames with thick and short x (red), y (green), and z (blue) orthogonal axes, and the estimates in the list returned by SLIDE as coordinate frames with thin and long axes. (a) Rotation search with 2 rotations, each generating 50% of the measurements. (b) Rotation search with 5 rotations, each generating 20% of the measurements.

9 Extensions and Open Problems

Extensions to other geometric perception problems. In this monograph we focused on

problems where inliers can be expressed via a linear model with additive noise: $\boldsymbol{y}_i = \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ + \boldsymbol{\epsilon}$. However, in many geometric perception problems, the measurements do not belong to a vector space and the noise is no longer additive. In the following, we review three examples of such problems, and observe that —in the outlier-free case— they can still be cast as:

$$\min_{\boldsymbol{x} \in \mathbb{X}} \ \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x} \right\|_2^2, \tag{50}$$

hence still being amenable for estimators such as (LTS), (MC), and (TLS) with $f_i(\boldsymbol{x}) = \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}$.

Example 18 (Single Rotation Averaging). *Estimate a rotation $\boldsymbol{R} \in \mathrm{SO}(3)$ given rotation measurements $\boldsymbol{R}_i$, $i = 1, \ldots, n$. The measurement model for the (inlier) measurements is given by:*

$$\boldsymbol{R}_i = \boldsymbol{R} \cdot \boldsymbol{R}_\epsilon, \qquad i \in [n], \tag{51}$$

where $\boldsymbol{R}_\epsilon$ is a random rotation describing the measurement noise. When the rotation noise $\boldsymbol{R}_\epsilon$ follows an isotropic Langevin distribution with mode $\mathbf{I}_3$ and concentration parameter κ (inconsequential in this example), a maximum likelihood estimator for the outlier-free single rotation averaging problem is given by the following optimization problem (where $\|\cdot\|_\mathsf{F}$ denotes the Frobenius norm):

$$\min_{\boldsymbol{R} \in \mathrm{SO}(3)} \sum_{i=1}^{n} \kappa \left\| \boldsymbol{R} - \boldsymbol{R}_i \right\|_\mathsf{F}^2, \tag{52}$$

which has the same form of (50) after vectorizing the matrix into a 9-vector. Rotation averaging finds application in camera calibration, motion capture, spacecraft attitude determination, and crystallography, see [121, 4] and the references therein.

Example 19 (Multiple Rotation Averaging). *Estimate a set of rotations $\boldsymbol{R}_k \in \mathrm{SO}(3)$, $k = 1, \ldots, N$, from relative rotation measurements $\bar{\boldsymbol{R}}_{ij}$ between (a sufficiently large set of) pairs of rotations. The generative model for the (inlier) measurements is:*

$$\bar{\boldsymbol{R}}_{ij} = \boldsymbol{R}_i^\mathsf{T} \boldsymbol{R}_j \boldsymbol{R}_\epsilon, \qquad (i, j) \in \mathcal{E}, \tag{53}$$

where $\mathcal{E}$ is the set of pairs (i, j) such that a measurement $\bar{\boldsymbol{R}}_{ij}$ is available. The problem can be visualized as a graph, where each node is associated a to-be-estimated rotation, while edges correspond to pairwise rotation measurements. When the rotation noise $\boldsymbol{R}_\epsilon$ follows an isotropic Langevin distribution with mode $\mathbf{I}_3$ and concentration parameter κ (inconsequential also in this example), a maximum likelihood estimator for the outlier-free multiple rotation averaging problem is given by the following optimization problem [122]:

$$\min_{\boldsymbol{R}_k \in \mathrm{SO}(3), k=1,\ldots,N} \sum_{(i,j) \in \mathcal{E}} \kappa \left\| \boldsymbol{R}_j - \boldsymbol{R}_i \bar{\boldsymbol{R}}_{ij} \right\|_\mathsf{F}^2, \tag{54}$$

which has the same form of (50) after vectorization. Multiple rotation averaging arises in Structure from Motion and camera calibration [4] among other problems, and can be used to compute an initial guess for Simultaneous Localization and Mapping methods [123]. It has also been studied in conjunctions with Shor's relaxation in [46, 122].

Example 20 (Pose Graph Optimization). *Estimate a set of poses $(\boldsymbol{t}_k, \boldsymbol{R}_k)$ with $\boldsymbol{t}_k \in \mathbb{R}^3$ and $\boldsymbol{R}_k \in \mathrm{SO}(3)$, $k = 1, \ldots, N$, from relative pose measurements $(\bar{\boldsymbol{t}}_{ij}, \bar{\boldsymbol{R}}_{ij})$ between (a sufficiently large set of) pairs of poses. The generative model for the (inlier) measurements is:*

$$\bar{\boldsymbol{R}}_{ij} = \boldsymbol{R}_i^\mathsf{T} \boldsymbol{R}_j \boldsymbol{R}_\epsilon, \qquad \bar{\boldsymbol{t}}_{ij} = \boldsymbol{R}_i^\mathsf{T} (\boldsymbol{t}_j - \boldsymbol{t}_i) + \boldsymbol{t}_\epsilon, \qquad (i,j) \in \mathcal{E}, \tag{55}$$

where $\mathcal{E}$ is the set of pairs (i,j) such that a measurement $(\bar{\boldsymbol{t}}_{ij}, \bar{\boldsymbol{R}}_{ij})$ is available. The problem can again be visualized as a graph, the pose graph*, where each node is associated a to-be-estimated pose, while edges correspond to pairwise pose measurements. When the rotation noise $\boldsymbol{R}_\epsilon$ follows an isotropic Langevin distribution with mode $\mathbf{I}_3$ and concentration parameter κ, and the translation error $\boldsymbol{t}_\epsilon$ is a zero-mean Gaussian with covariance $\frac{1}{\tau} \mathbf{I}_3$, a maximum likelihood estimator for the outlier-free pose graph optimization problem is given by the following optimization problem [13, 110, 7]:*

$$\min_{\boldsymbol{t}_k \in \mathbb{R}^3, \boldsymbol{R}_k \in \mathrm{SO}(3), k=1,\ldots,N} \sum_{(i,j) \in \mathcal{E}} \tau \left\| \boldsymbol{t}_j - \boldsymbol{t}_i - \boldsymbol{R}_i \bar{\boldsymbol{t}}_{ij} \right\|_2^2 + \sum_{(i,j) \in \mathcal{E}} \kappa \left\| \boldsymbol{R}_j - \boldsymbol{R}_i \bar{\boldsymbol{R}}_{ij} \right\|_\mathsf{F}^2, \tag{56}$$

which, observing the quadratic nature of the cost in (56)*, can be recast as in eq.* (50)*. Pose graph optimization finds application in Simultaneous Localization and Mapping among other fields [5] and has been investigated in conjunction with Shor's relaxation in [13, 110, 7, 124, 14, 125].*

All the examples above still reduce to linear regression over a basic semi-algebraic set, hence we believe it is possible to extend the results presented in this monograph to these problems. Indeed, moment relaxations of a (TLS) formulation of outlier-robust pose graph optimization have been proposed in [35]. At the same time, multiple rotation averaging and pose graph optimization pose further challenges, due to the very high-dimensional nature of the problem (for which even a sparse moment relaxation of order 2 is out of reach for current SDP solvers and for realistic problem sizes [35]), and might benefit from different assumptions on the measurements that better leverage the graph-theoretic nature of the problem (*e.g.*, problems (53) and (56) admit a unique solution only if the underlying graph is connected).

Other extensions and open problems. This monograph extends results from outlier-robust statistics to problems with vector-valued measurements, unknown outlier rates, and variables belonging to a basic semi-algebraic set $\mathbb{X}$. However, all the algorithms presented in this monograph may return estimates outside $\mathbb{X}$. While in principle, for the sets arising in geometric perception, it is typically easy to project the outputs of these algorithms onto $\mathbb{X}$, the corresponding estimation contracts currently do not account for such a rounding. Leveraging Lemma 3 in [126], we could account for the rounding by adding a constant factor of 2 in front of the error bounds in Propositions 13 to 15 and Theorem 17.[33] However, it would be desirable to take advantage of the domain constraint to *tighten* the error bounds in the proofs, rather than treating it as an afterthought (in this monograph, we mostly make sure that the domain constraint does not break the proofs).

A second limitation is that the estimation contracts presented in this monograph require the corresponding algorithms to solve high-order moment relaxations. For instance, we have already observed that Theorem 12 requires a relaxation of order $r \geq k \geq 4$ (under k-certifiable hypercontractivity), which would be impractical to solve with current SDP solvers; in this case the bound $r \geq k \geq 4$ on the relaxation order is imposed by the degree of the objective in Algorithm 1 (which imposes $r \geq k$), and by the proof (which requires $k \geq 4$). This requirement is further reinforced by the fact that typical distributions (*e.g.*, Gaussians) have been shown to be 4-certifiably hypercontractive [15],

---

[33]We thank Kevin Doherty for pointing out the result in [126].

hence again requiring $k \geq 4$ for the performance guarantees to hold. The same issue arises for the other estimation contracts, due to the requirement that the relaxation order must satisfy $r \geq k/2$ for k-certifiably anti-concentrated inliers and that $k \geq 6$, due to the degree of the polynomials in (33). Therefore, it would be desirable to develop guarantees for more practical (and possibly sparse) order-2 relaxations, with the goal of fully explaining the empirical performance observed in [30].

Third, it would be interesting to explore the performance of list-decodable estimation in other perception problems. In particular, it would be useful to see if the impressive performance observed in rotation search problems (Section 8.5) can be also replicated in other geometric perception problems. Along these lines, it would be desirable to develop more sophisticated algorithmic variants of SLIDE based on the framework in [37]. For instance, rather than sampling potential estimates from the pseudo-moment matrix, one could collect all estimates (as done in SLIDE), cluster them into a smaller number of hypotheses, and return the cluster centers, hoping that the cluster-wise averaging can further reduce the estimation errors. Related to the discussion above, it would be desirable to extend the performance guarantees in Theorem 17 to encompass order-2 relaxations and capture the fact that projecting the estimates to the domain $\mathbb{X}$ leads to much better performance in practice.

Fourth, it would be interesting to characterize the behavior of adversarial outliers in geometric perception. How does an adversarial attack on a geometric perception problem look like? While in the experiments we observed that "mutually consistent" outliers are more challenging to reject, a strong adversary model would also leverage the set of inliers to create even more difficult problem instances. Along these lines, designing algorithms to produce adversarial outliers in perception problems might provide further insights on outlier-robust estimation algorithms, somewhat drawing inspiration from the large and parallel body of research on adversarial machine learning, see, *e.g.,* [127].[34] A somewhat related problem is how to train the front-end of a perception system to *avoid* these adversarial instances. In other words, can we design a front-end (*i.e.,* train a neural-network) in a way that the resulting measurements satisfy our estimation contracts by construction? For instance, this would entail enforcing or encouraging certifiable hypercontractivity or certifiable anti-concentration of the measurements produced by the front-end during training.

Finally, a broader issue is that many of the estimation problems considered in this monograph must be solved on a stringent runtime budget. For instance, point cloud registration problems (Example 2) are solved at frame-rate (*e.g.,* > 20 Hz) in many RGB-D SLAM applications [5], hence requiring the estimation algorithm to run in a fraction of a second. SDP solvers applied to the moment relaxations considered in this monograph (even at order 2) are far from meeting this runtime constraints. Therefore, it would be interesting to develop specialized solvers that take advantage of the problem structure; for instance, the work [119] leverages the fact that the SDP is a relaxation of a polynomial optimization problem to speed up computation, while [7] achieves real-time performance by solving large SDPs using the Riemannian staircase method [129]. Finally, it is important to further extend the reach of sparse versions of Lasserre's hierarchy of moment relaxations, which can reduce the size of the matrices in the relaxation by leveraging the problem structure, see [130, 131, 132, 133].

10   Conclusions

We studied outlier-robust estimation in the context of the geometric perception problems arising in robotics and computer vision. Many of these problems can be reformulated as linear estimation

---

[34]An empirical study on adversarial attacks on SLAM has been recently proposed in [128].

problems with variables belonging to a basic semi-algebraic set, and the goal is to retrieve a good estimate of the variables in the presence of outliers. We provided a unified view of converging work on outlier-robust estimation across robust statistics, robotics, and computer vision and discussed technical tools underlying modern estimation approaches, including moment relaxations and sum-of-squares proofs. Then, we reviewed existing algorithms and presented *estimation contracts*, which establish conditions on the input measurements under which modern estimation algorithms are guaranteed to recover an estimate close to the ground truth in the presence of outliers. Towards this goal, we adapted and extended recent results on outlier-robust *linear regression* (applicable to the low-outlier case with $\ll 50\%$ outliers) and *list-decodable regression* (applicable to the high-outlier case with $\gg 50\%$ outliers) to the setup commonly found in robotics and vision, where (i) variables (*e.g.,* rotations, poses) belong to non-convex sets, (ii) measurements are vector-valued, and (iii) the number of outliers is not known a priori. Besides the technical results, we hope this monograph can provide a unifying view of parallel research lines on outlier-robust estimation across fields. Moreover, we hope that practitioners will benefit from our layman introduction to moment relaxations and sum-of-squares proofs and will use these tools to attack other outstanding problems in robotics and vision. Finally, we hope that researchers in robust statistics will be intrigued by the formulations and the empirical performance observed in robotics and vision problems, and will contribute to bridging the current gap between theoretical results and practical algorithms.

Acknowledgments

References

[1] H. Yang, J. Shi, and L. Carlone, "TEASER: Fast and Certifiable Point Cloud Registration," *IEEE Trans. Robotics*, vol. 37, no. 2, pp. 314–333, 2020, extended arXiv version 2001.07715 (pdf).

[2] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, "3D shape reconstruction from 2D landmarks: A convex formulation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] D. Scaramuzza and F. Fraundorfer, "Visual odometry: Part I the first 30 years and fundamentals," 2011.

[4] R. Hartley, J. Trumpf, Y. Dai, and H. Li, "Rotation averaging," *IJCV*, vol. 103, no. 3, pp. 267–305, 2013.

[5] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-

perception age," *IEEE Trans. Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016, arxiv preprint: 1606.05830, (pdf).

[6] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *International workshop on vision algorithms.* Springer, 1999, pp. 298–372.

[7] D. Rosen, L. Carlone, A. Bandeira, and J. Leonard, "SE-Sync: a certifiably correct algorithm for synchronization over the Special Euclidean group," *Intl. J. of Robotics Research*, 2018, arxiv preprint: 1611.00128, (pdf).

[8] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *J. Opt. Soc. Amer.*, vol. 4, no. 4, pp. 629–642, Apr 1987.

[9] F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Foundations and Trends in Robotics*, vol. 6, no. 1-2, pp. 1–139, 2017.

[10] V. Larsson, M. Oskarsson, K. Astrom, A. Wallis, T. Pajdla, and Z. Kukelova, "Beyond Grobner bases: Basis selection for minimal solvers," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3945–3954.

[11] J. Shi, H. Yang, and L. Carlone, "Optimal pose and shape estimation for category-level 3D object perception," in *Robotics: Science and Systems (RSS)*, 2021, arXiv preprint: 2104.08383, (pdf), (video).

[12] F. Kahl and D. Henrion, "Globally optimal estimates for geometric reconstruction problems," *Intl. J. of Computer Vision*, vol. 74, no. 1, pp. 3–15, 2007.

[13] L. Carlone and F. Dellaert, "Duality-based verification techniques for 2D SLAM," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 4589–4596, (pdf) (code).

[14] L. Carlone, G. Calafiore, C. Tommolillo, and F. Dellaert, "Planar pose graph optimization: Duality, optimal solutions, and verification," *IEEE Trans. Robotics*, vol. 32, no. 3, pp. 545–565, 2016, (pdf) (code).

[15] A. R. Klivans, P. K. Kothari, and R. Meka, "Efficient algorithms for outlier-robust regression," *CoRR*, vol. abs/1803.03241, 2018. [Online]. Available: http://arxiv.org/abs/1803.03241

[16] J. W. Tukey, "Mathematics and the picturing of data," *Proceedings of the International Congress of Mathematicians*, vol. 2, pp. 523–531, 1975. [Online]. Available: https://cir.nii.ac.jp/crid/1573950399770196096

[17] P. Huber, *Robust Statistics.* John Wiley & Sons, New York, NY, 1981.

[18] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, NY, 1981.

[19] ——, *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, NY, 1987.

[20] T. Bernholt, "Robust estimators are hard to compute," Dortmund, Technical Report 2005,52, 2006. [Online]. Available: http://hdl.handle.net/10419/22645

[21] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart, "Sever: A robust meta-algorithm for stochastic optimization," in *Intl. Conf. on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 2019, pp. 1596–1606.

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

[23] I. Diakonikolas, W. Kong, and A. Stewart, "Efficient algorithms and lower bounds for robust linear regression," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '19, 2019, pp. 2745–2754.

[24] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar, "Consistent robust regression," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30. Curran Associates, Inc., 2017.

[25] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM J. Optim.*, vol. 11, no. 3, pp. 796–817, 2001.

[26] P. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, California Institute of Technology, 2000.

[27] T. J. Chin and D. Suter, "The maximum consensus problem: recent algorithmic advances," *Synthesis Lectures on Computer Vision*, vol. 7, no. 2, pp. 1–194, 2017.

[28] P. Antonante, V. Tzoumas, H. Yang, and L. Carlone, "Outlier-robust estimation: Hardness, minimally tuned algorithms, and applications," *IEEE Trans. Robotics*, vol. 38, no. 1, pp. 281–301, 2021, (pdf).

[29] T.-J. Chin, Z. Cai, and F. Neumann, "Robust fitting in computer vision: Easy or hard?" in *European Conf. on Computer Vision (ECCV)*, 2018.

[30] H. Yang and L. Carlone, "Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, 2022, (pdf).

[31] K. M. Tavish and T. D. Barfoot, "At all costs: A comparison of robust cost functions for camera correspondence outliers," in *Conf. Computer and Robot Vision*. IEEE, 2015, pp. 62–69.

[32] H. Yang, P. Antonante, V. Tzoumas, and L. Carlone, "Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 1127–1134, 2020, arXiv preprint:1909.08605 (with supplemental material), (pdf).

[33] H. Yang and L. Carlone, "One ring to rule them all: Certifiably robust geometric perception with outliers," in *Conf. on Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18846–18859, (pdf). [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/da6ea77475918a3d83c7e49223d453cc-Paper.pdf

[34] ——, "A quaternion-based certifiably optimal solution to the Wahba problem with outliers," in *Intl. Conf. on Computer Vision (ICCV)*, 2019, (Oral Presentation, accept rate: 4%), Arxiv version: 1905.12536, (pdf).

[35] P. Lajoie, S. Hu, G. Beltrame, and L. Carlone, "Modeling perceptual aliasing in SLAM via discrete-continuous graphical models," *IEEE Robotics and Automation Letters (RA-L)*, 2019, extended ArXiv version: (pdf), Supplemental Material: (pdf).

[36] L. Peng, M. Fazlyab, and R. Vidal, "Towards understanding the semidefinite relaxations of truncated least-squares in robust rotation search," 2022. [Online]. Available: https://arxiv.org/abs/2207.08350

[37] S. Karmalkar, A. Klivans, and P. Kothari, "List-decodable linear regression," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 32, 2019.

[38] L. Carlone and G. Calafiore, "Convex relaxations for pose graph optimization with outliers," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 2, pp. 1160–1167, 2018, arxiv preprint: 1801.02112, (pdf).

[39] N. Shor, "Nondifferentiable optimization and polynomial problems," *Nonconvex Optimization and its Applications*, vol. 24, 1998.

[40] Y. Nesterov, "Squared functional systems and optimization problems, high performance optimization," *Appl. Optim.*, vol. 33, pp. 405–440, 2000.

[41] P. A. Parrilo, "Semidefinite programming relaxations for semialgebraic problems," *Mathematical programming*, vol. 96, no. 2, pp. 293–320, 2003.

[42] Y. Bar-Shalom and X. Li, *Estimation and Tracking: principles, techniques and software*. Boston, London: Artech House, 1993.

[43] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte Carlo Localization for mobile robots," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 1999.

[44] M. Hsiao and M. Kaess, "MH-iSAM2: multi-hypothesis iSAM using Bayes Tree and Hypo-tree," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2019, pp. 1274–1280.

[45] D. Fourie, J. Leonard, and M. Kaess, "A nonparametric belief solution to the bayes tree," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2016, pp. 2189–2196.

[46] A. Eriksson, C. Olsson, F. Kahl, and T.-J. Chin, "Rotation averaging and strong duality," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[47] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[48] G. Blekherman, P. A. Parrilo, and R. R. Thomas, *Semidefinite optimization and convex algebraic geometry*. SIAM, 2012.

[49] B. Barak and D. Steurer, "Proofs, beliefs, and algorithms through the lens of sum-of-squares," in *Lecture notes in preparation, available on http://sumofsquares.org*, 2016.

[50] N. Fleming, P. Kothari, and T. Pitassi, "Semialgebraic proofs and efficient algorithm design," *Foundations and Trends in Theoretical Computer Science*, vol. 14, no. 1-2, pp. 1–221, 2019.

[51] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.

[52] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *Intl. J. of Computer Vision*, vol. 19, no. 1, pp. 57–91, 1996.

[53] Á. Parra Bustos and T. J. Chin, "Guaranteed outlier removal for rotation search," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2165–2173.

[54] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4331–4339.

[55] N. Chebrolu, T. Läbe, O. Vysotska, J. Behley, and C. Stachniss, "Adaptive robust kernels for non-linear least squares problems," *arXiv preprint arXiv:2004.14938*, 2020.

[56] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[57] P. Agarwal, G. D. Tipaldi, L. Spinello, C. Stachniss, and W. Burgard, "Robust map optimization using dynamic covariance scaling," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2013.

[58] N. Sunderhauf and P. Protzel, "Towards a robust back-end for pose graph SLAM," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2012, pp. 1254–1261.

[59] T. Hitchcox and J. R. Forbes, "Mind the gap: Norm-aware adaptive robust loss for multivariate least-squares problems," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7116–7123, 2022.

[60] H. M. Le, T.-J. Chin, A. Eriksson, T.-T. Do, and D. Suter, "Deterministic approximate methods for maximum consensus robust fitting," *IEEE Trans. Pattern Anal. Machine Intell.*, 2019.

[61] J. C. Bazin, Y. Seo, and M. Pollefeys, "Globally optimal consensus set maximization through rotation search," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 539–551.

[62] Á. Parra Bustos and T. J. Chin, "Guaranteed outlier removal for point cloud registration with correspondences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 40, no. 12, pp. 2868–2882, 2018.

[63] G. Izatt, H. Dai, and R. Tedrake, "Globally optimal object pose estimation in point clouds with mixed-integer programming," in *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, 2017.

[64] J. Yang, H. Li, and Y. Jia, "Optimal essential matrix estimation via inlier-set maximization," in *European Conf. on Computer Vision (ECCV)*. Springer, 2014, pp. 111–126.

[65] D. P. Paudel, A. Habed, C. Demonceaux, and P. Vasseur, "Robust and optimal sum-of-squares-based point-to-plane registration of image sets and structured scenes," in *Intl. Conf. on Computer Vision (ICCV)*, 2015, pp. 2048–2056.

[66] H. Li, "Consensus set maximization with guaranteed global optimality for robust geometry estimation," in *Intl. Conf. on Computer Vision (ICCV)*, 2009, pp. 1074–1080.

[67] H. Li and R. Hartley, "The 3D-3D registration problem revisited," in *Intl. Conf. on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[68] O. Enqvist and F. Kahl, "Robust optimal pose estimation," in *European Conf. on Computer Vision (ECCV)*. Springer, 2008, pp. 141–153.

[69] O. Enqvist, E. Ask, F. Kahl, and K. Åström, "Tractable algorithms for robust model estimation," *Intl. J. of Computer Vision*, vol. 112, no. 1, pp. 115–129, 2015.

[70] ——, "Robust fitting for multiple view geometry," in *European Conf. on Computer Vision (ECCV)*. Springer, 2012, pp. 738–751.

[71] C. Olsson, O. Enqvist, and F. Kahl, "A polynomial-time bound for matching and registration with outliers," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[72] T.-J. Chin, P. Purkait, A. Eriksson, and D. Suter, "Efficient globally optimal consensus maximisation with tree search," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2413–2421.

[73] K. A. Lai, A. B. Rao, and S. Vempala, "Agnostic estimation of mean and covariance," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society, 2016, pp. 665–674. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/FOCS.2016.76

[74] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," in *IEEE 57th Annual Symposium on Foundations of Computer Science*. IEEE, 2016, pp. 655–664.

[75] ——, "Robust estimators in high-dimensions without the computational intractability," *SIAM Journal on Computing*, vol. 48, no. 2, pp. 742–864, 2019.

[76] M. Charikar, J. Steinhardt, and G. Valiant, "Learning from untrusted data," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2017, 2017, pp. 47–60.

[77] P. K. Kothari and J. Steinhardt, "Better agnostic clustering via relaxed tensor norms," *CoRR*, vol. abs/1711.07465, 2017. [Online]. Available: http://arxiv.org/abs/1711.07465

[78] P. K. Kothari, J. Steinhardt, and D. Steurer, "Robust moment estimation and improved clustering via sum of squares," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2018, 2018, pp. 1035–1046.

[79] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Robustly learning a gaussian: Getting optimal error, efficiently," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '18, 2018, p. 2683–2702.

[80] A. R. Klivans, P. M. Long, and R. A. Servedio, "Learning halfspaces with malicious noise," in *Automata, Languages and Programming*, S. Albers, A. Marchetti-Spaccamela, Y. Matias, S. Nikoletseas, and W. Thomas, Eds., 2009, pp. 609–621.

[81] I. Diakonikolas, D. M. Kane, and A. Stewart, "Learning geometric concepts with nasty noise," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2018, 2018, pp. 1061–1073.

[82] P. Awasthi, M. F. Balcan, and P. M. Long, "The power of localization for efficiently learning linear separators with noise," *J. ACM*, vol. 63, no. 6, 2017.

[83] S. Karmalkar and E. Price, "Compressed sensing with adversarial sparse noise via L1 regression," *CoRR*, vol. abs/1809.08055, 2018. [Online]. Available: http://arxiv.org/abs/1809.08055

[84] S. S. Du, S. Balakrishnan, and A. Singh, "Computationally efficient robust estimation of sparse functionals," *CoRR*, vol. abs/1702.07709, 2017. [Online]. Available: https://arxiv.org/abs/1702.07709

[85] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 721–729.

[86] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *Intl. Conf. on Machine Learning (ICML)*, vol. 28, no. 3, 2013, pp. 774–782.

[87] N. Nguyen and T. Tran, "Exact recoverability from dense corrupted observations via ℓ_1-minimization," *IEEE Trans. on Information Theory*, vol. 59, no. 4, pp. 2017–2035, 2013.

[88] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," *IEEE Trans. on Information Theory*, vol. 56, no. 7, pp. 3540–3560, 2010.

[89] T. D'Orsi, G. Novikov, and D. D. Steurer, "Consistent regression when oblivious outliers overwhelm," in *Intl. Conf. on Machine Learning (ICML)*, vol. 139, 2021, pp. 2297–2306.

[90] E. Tsakonas, J. Jaldén, N. Sidiropoulos, and B. Ottersten, "Convergence of the huber regression M-estimate in the presence of dense outliers," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1211–1214, 2014.

[91] T. Norman, N. Weinberger, and K. Levy, "Robust linear regression for general feature distribution," *ArXiv*, vol. abs/2202.02080, 2022.

[92] P. Raghavendra and M. Yau, "List decodable learning via sum of squares," in *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '20, 2020, p. 161–180.

[93] M.-F. Balcan, A. Blum, and S. Vempala, "A discriminative framework for clustering via similarity functions," in *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, ser. STOC '08, 2008, pp. 671–680.

[94] I. Diakonikolas, D. M. Kane, and A. Stewart, "List-decodable robust mean estimation and learning mixtures of spherical gaussians," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2018, 2018, pp. 1047–1060.

[95] R. Tron, D. Rosen, and L. Carlone, "On the inclusion of determinant constraints in lagrangian duality for 3D SLAM," in *Robotics: Science and Systems (RSS), Workshop "The problem of mobile sensors: Setting future goals and indicators of progress for SLAM"*, 2015, (pdf).

[96] J. Briales, L. Kneip, and J. Gonzalez-Jimenez, "A certifiably globally optimal solution to the non-minimal relative pose problem," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[97] H. Yang and L. Carlone, "In perfect shape: Certifiably optimal 3D shape reconstruction from 2D landmarks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, arxiv version: 1911.11924, (pdf).

[98] G. Wahba, "A least squares estimate of satellite attitude," *SIAM review*, vol. 7, no. 3, pp. 409–409, 1965.

[99] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, January 1995.

[100] J. Shi, H. Yang, and L. Carlone, "Optimal and robust category-level perception: Object pose and shape estimation from 2D and 3D semantic keypoints," *arXiv preprint: 2206.12498*, 2022, (pdf).

[101] L. Kneip, H. Li, and Y. Seo, "UPnP: An optimal o(n) solution to the absolute pose problem with universal applicability," in *European Conf. on Computer Vision (ECCV)*. Springer, 2014, pp. 127–142.

[102] G. Schweighofer and A. Pinz, "Globally optimal O(n) solution to the PnP problem for general camera models." in *British Machine Vision Conf. (BMVC)*, 2008, pp. 1–10.

[103] H. Yang, C. Doran, and J.-J. Slotine, "Dynamical pose estimation," in *Intl. Conf. on Computer Vision (ICCV)*, 2021.

[104] J. Lasserre, *Moments, positive polynomials and their applications.* World Scientific, 2010, vol. 1.

[105] MOSEK ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 8.1.*, 2017. [Online]. Available: http://docs.mosek.com/8.1/toolbox/index.html

[106] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming." [Online]. Available: http://cvxr.com/cvx

[107] R. H. Tütüncü, K.-C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical programming*, vol. 95, no. 2, pp. 189–217, 2003.

[108] J. Lasserre, "The Moment-SoS hierarchy," in *Int. Cong. of Math.*, vol. 4, 2018, pp. 3791–3815.

[109] J. Nie, "Optimality conditions and finite convergence of lasserre's hierarchy," *Mathematical programming*, vol. 146, no. 1-2, pp. 97–121, 2014.

[110] L. Carlone, D. Rosen, G. Calafiore, J. Leonard, and F. Dellaert, "Lagrangian duality in 3D SLAM: Verification techniques and optimal solutions," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 125–132, (pdf) (code) (datasets: (web)) (supplemental material: (pdf)).

[111] J. Briales and J. Gonzalez-Jimenez, "Convex Global 3D Registration with Lagrangian Duality," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[112] S. B. Hopkins and J. Li, "Mixture models, robustness, and sum of squares proofs," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2018, 2018, pp. 1021–1034.

[113] T. Ma, J. Shi, and D. Steurer, "Polynomial-time tensor decompositions with sum-of-squares," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 2016, pp. 438–446.

[114] I. Diakonikolas, D. M. Kane, S. Karmalkar, A. Pensia, and T. Pittas, "Robust sparse mean estimation via sum of squares," in *Proceedings of Thirty Fifth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 178. PMLR, 2022, pp. 4703–4763.

[115] B. Barak, F. Brandao, A. Harrow, J. Kelner, D. Steurer, and Y. Zhou, "Hypercontractivity, sum-of-squares proofs, and their applications," in *Proc. of the Annual ACM Symp. on Theory of Computing*, 2012, pp. 307–326.

[116] A. Bakshi and P. K. Kothari, "List-decodable subspace recovery: Dimension independent error in polynomial time," in *Proc. of the Annual ACM-SIAM Symp. on Discrete Algorithms (SODA)*, ser. SODA '21, 2021, pp. 1279–1297.

[117] G. Meneghetti, M. Danelljan, M. Felsberg, and K. Nordberg, "Image alignment for panorama stitching in sparsely structured environments," in *Scandinavian Conference on Image Analysis*. Springer, 2015, pp. 428–439.

[118] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: speeded up robust features," in *European Conf. on Computer Vision (ECCV)*, 2006.

[119] H. Yang, L. Liang, L. Carlone, and K. Toh, "An inexact projected gradient method with rounding and lifting by nonlinear programming for solving rank-one semidefinite relaxation of polynomial optimization," *Mathematical Programming (MAPR)*, 2022, (pdf).

[120] A. Papachristodoulou, J. Anderson, G. Valmorbida, S. Prajna, P. Seiler, and P. A. Parrilo, *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, http://arxiv.org/abs/1310.4716, 2013.

[121] S. H. Lee and J. Civera, "Robust single rotation averaging," *CoRR*, vol. abs/2004.00732, 2020. [Online]. Available: https://arxiv.org/abs/2004.00732

[122] F. Dellaert, D. Rosen, J. Wu, R. Mahony, and L. Carlone, "Shonan rotation averaging: Global optimality by surfing $SO(p)^n$," in *European Conf. on Computer Vision (ECCV)*, 2020, accepted as spotlight presentation (acceptance rate 5%) (pdf), (web).

[123] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert, "Initialization techniques for 3D SLAM: a survey on rotation estimation and its use in pose graph optimization," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 4597–4604, (pdf) (code) (supplemental material: (pdf)).

[124] J. Briales and J. Gonzalez-Jimenez, "Cartan-sync: Fast and global SE(d)-synchronization," *IEEE Robot. Autom. Lett*, vol. 2, no. 4, pp. 2127–2134, 2017.

[125] T. Fan, H. Wang, M. Rubenstein, and T. Murphey, "Efficient and guaranteed planar pose graph optimization using the complex number representation," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 1904–1911.

[126] K. Doherty, D. M. Rosen, and J. Leonard, "Performance guarantees for spectral initialization in rotation averaging and pose-graph SLAM," *CoRR*, vol. abs/2201.03773, 2022. [Online]. Available: https://arxiv.org/abs/2201.03773

[127] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv e-print: 1412.6572*, 2014.

[128] M. Ikram, S. Khaliq, M. Anjum, and W. Hussain, "Perceptual aliasing++: Adversarial attack for visual SLAM front-end and back-end," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4670–4677, 2022.

[129] N. Boumal, V. Voroninski, and A. Bandeira, "The non-convex Burer-Monteiro approach works on smooth semidefinite programs," *arXiv*, 2016.

[130] T. Weisser, J. B. Lasserre, and K.-C. Toh, "Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity," *Math. Program. Comput.*, vol. 10, no. 1, pp. 1–32, 2018.

[131] J. Wang, V. Magron, and J.-B. Lasserre, "Chordal-TSSOS: a moment-SOS hierarchy that exploits term sparsity with chordal extension," *SIAM Journal on Optimization*, vol. 31, no. 1, pp. 114–141, 2021.

[132] J. Wang, V. Magron, J. B. Lasserre, and N. H. A. Mai, "CS-TSSOS: Correlative and term sparsity for large-scale polynomial optimization," *arXiv preprint arXiv:2005.02828*, 2020.

[133] J. Wang, V. Magron, and J.-B. Lasserre, "TSSOS: A Moment-SOS hierarchy that exploits term sparsity," *SIAM Journal on Optimization*, vol. 31, no. 1, pp. 30–58, 2021.

[134] B. Barak, J. A. Kelner, and D. Steurer, "Rounding sum-of-squares relaxations," in *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, ser. STOC '14, 2014, pp. 31–40.

A An Algorithmic View of Lasserre's Hierarchy of Moment Relaxations

Here we provide an algorithmic (and somewhat unorthodox) view of Lasserre's hierarchy of moment relaxations [25]; we refer the reader to [108, 104] for a more standard introduction.

Lasserre's hierarchy provides a systematic way to relax a polynomial optimization problem (POP) into a semidefinite (convex) program. We start by restating (POP):

$$p^\star \triangleq \min_{\boldsymbol{x} \in \mathbb{R}^{d_x}} \left\{ p(\boldsymbol{x}) \,\middle|\, \begin{array}{l} h_i(\boldsymbol{x}) = 0, i = 1, \ldots, l_h \\ g_j(\boldsymbol{x}) \geq 0, j = 1, \ldots, l_g \end{array} \right\}, \tag{POP}$$

where $p(\boldsymbol{x}), h_i(\boldsymbol{x}), g_j(\boldsymbol{x})$ are polynomials in the variable $\boldsymbol{x} \in \mathbb{R}^{d_x}$.

The key idea behind Lasserre's hierarchy of moment relaxations is to (i) rewrite the polynomial optimization problem (POP) using the moment matrix $\boldsymbol{X}_{2r}$[35], (ii) relax the (non-convex) rank-1 constraint on $\boldsymbol{X}_{2r}$, and (iii) add redundant constraints that are trivially satisfied in (POP) but might still improve the quality of the relaxation; as shown below, this leads to a semidefinite program.

(i) Rewriting (POP) using $\boldsymbol{X}_{2r}$. Recall that any polynomial of degree up to $2r$ can be written as a linear function of the moment matrix $\boldsymbol{X}_{2r}$ (*cf.* Section 4.2). Therefore, we pick a positive integer r (the *order* of the relaxation) such that $2r \geq \max\{\deg(p), \deg(h_1), \ldots, \deg(h_{l_h}), \deg(g_1), \ldots, \deg(g_{l_g})\}$, such that we can express both objective function and constraints as a linear function of $\boldsymbol{X}_{2r}$. With this choice of r, we can rewrite the objective and the equality constraints in (POP) as:

$$\text{objective}: \quad \langle \boldsymbol{C}_1, \boldsymbol{X}_{2r} \rangle \tag{A1}$$

$$\text{equality constraints}: \langle \boldsymbol{A}_{\text{eq},j}, \boldsymbol{X}_{2r} \rangle = 0, \; j = 1, \ldots, l_h, \tag{A2}$$

for suitable matrices $\boldsymbol{C}_1$ and $\boldsymbol{A}_{\text{eq},j}$.

(ii) Relaxing the (non-convex) rank-1 constraint on $\boldsymbol{X}_{2r}$. At the previous point we noticed we can rewrite objective and constraints in (POP) as linear (hence convex) functions of $\boldsymbol{X}_{2r}$. However, $\boldsymbol{X}_{2r}$ still belongs to the set of positive-semidefinite rank-1 matrices (since it is defined as $[\boldsymbol{x}]_r [\boldsymbol{x}]_r^\mathsf{T}$, where $[\boldsymbol{x}]_r$ is a vector of monomials), which is a non-convex set due to the rank constraint. Therefore, we simply relax the rank constraint and only enforce:

$$\text{pseudo-moment matrix}: \quad \boldsymbol{X}_{2r} \succeq 0. \tag{A3}$$

(iii) Adding redundant constraints. Since we have relaxed (POP) by re-parametrizing it using $\boldsymbol{X}_{2r}$ and dropping the rank constraint, the final step to obtain Lasserre's relaxation consists in adding extra constraints to make the relaxation tighter. First of all, we observe that there are multiple repeated entries in the moment matrix (*e.g.*, in (9), the entry $x_1 x_2$ appears 4 times in the matrix). Therefore, we can enforce these entries to be the same. In general, this leads to $m_{\text{mom}} = \mathfrak{t}(\underline{d}_r) - \underline{d}_{2r} + 1$ linear constraints, where $\underline{d}_{2r} \triangleq \binom{d_x + 2r}{2r}$ (the size of the monomial basis of degree up to $2r$, *i.e.*, $[\boldsymbol{x}]_{2r}$) and $\mathfrak{t}(n) \triangleq \frac{n(n+1)}{2}$ is the dimension of $\mathbb{S}^n$. These constraints are typically called *moment constraints*:

$$\begin{aligned}
\text{moment constraints}: \quad & \langle \boldsymbol{A}_{\text{mom},0}, \boldsymbol{X}_{2r} \rangle = 1, \\
& \langle \boldsymbol{A}_{\text{mom},j}, \boldsymbol{X}_{2r} \rangle = 0, \\
& j = 1, \ldots, \mathfrak{t}(\underline{d}_r) - \underline{d}_{2r},
\end{aligned} \tag{A4}$$

---

[35] Recall that the moment matrix is defined as $\boldsymbol{X}_{2r} \triangleq [\boldsymbol{x}]_r [\boldsymbol{x}]_r^\mathsf{T}$, where $[\boldsymbol{x}]_r$ is the vector of monomials of degree up to r. For instance, for $\boldsymbol{x} = [x_1 \,;\, x_2]$ and $r = 2$, the matrix $\boldsymbol{X}_{2r}$ takes the form in eq. (9).

where $\boldsymbol{A}_{\mathrm{mom},0}$ is all-zero except $[\boldsymbol{A}_{\mathrm{mom},0}]_{11} = 1$, and it is used to define the constraint $[\boldsymbol{X}_{2r}]_{11} = 1$, following from the definition of the moment matrix (see eq. (9)).

Second, we can also add *redundant* equality constraints. Simply put, if $h_i = 0$, then also $h_i \cdot x_1 = 0$, $h_i \cdot x_2 = 0$, and so on, for any monomial we multiply by h_i. Since via $\boldsymbol{X}_{2r}$ we can represent any polynomial of degree up to $2r$, we can write as linear constraints any polynomial equality in the form $h_i \cdot [\boldsymbol{x}]_{2r-\deg(h_i)} = \boldsymbol{0}$ (the degree of the monomials is chosen such that the product does not exceed degree $2r$). These new equalities can again be written linearly as:

$$\text{(redundant) equality constraints} : \langle \boldsymbol{A}_{\mathrm{req},ij}, \boldsymbol{X}_{2r} \rangle = 0,$$
$$i = 1, \ldots, l_h, \quad j = 1, \ldots, \underline{d}_{2r-\deg(h_i)}, \tag{A5}$$

for suitable $\boldsymbol{A}_{\mathrm{req},ij}$. Since the first entry of $[\boldsymbol{x}]_{2r-\deg(h_i)}$ is always 1 (*i.e.*, the monomial of order zero), eq. (A5) already includes the original equality constraints in (A2).

Finally, we observe that if $g_j \geq 0$, then for any positive semidefinite matrix $\boldsymbol{M}$, it holds $g_j \cdot \boldsymbol{M} \succeq 0$. Since we can represent any polynomial of order up to $2r$ as a linear function of $\boldsymbol{X}_{2r}$, we can add redundant constraints in the form $g_j \cdot \boldsymbol{X}_{2(r-\lceil \deg(g_j)/2 \rceil)} \succeq 0$ (by construction $g_j \cdot \boldsymbol{X}_{2(r-\lceil \deg(g_j)/2 \rceil)}$ only contains polynomials of degree up to $2r$). To phrase the resulting relaxation in the standard form (SDP), it is common to add extra matrix variables $\boldsymbol{X}_{g_j} = g_j \cdot \boldsymbol{X}_{2(r-\lceil \deg(g_j)/2 \rceil)}$ for $j = 1, \ldots, l_g$ (the *localizing matrices* [104, §3.2.1]) and then force these matrices to be a linear function of $\boldsymbol{X}_{2r}$:

$$\text{localizing matrices} : \quad \boldsymbol{X}_{g_j} \succeq 0, \quad j = 1, \ldots, l_g, \tag{A6}$$

$$\text{localizing constraints} : \langle \boldsymbol{A}_{\mathrm{loc},jkh}, \boldsymbol{X}_{2r} \rangle = [\boldsymbol{X}_{g_j}]_{hk},$$
$$j = 1, \ldots, l_g, \quad 1 \leq h \leq k \leq \underline{d}_{r-\lceil \deg(g_j)/2 \rceil}, \tag{A7}$$

where the linear constraints (for some matrix $\boldsymbol{A}_{\mathrm{loc},jkh}$) enforce each entry of $\boldsymbol{X}_{g_j}$ to be a linear combination of entries of the matrix $\boldsymbol{X}_{2r}$.

Following steps (i)-(iii) above, it is straightforward to obtain the following semidefinite program:

$$f_{2r}^{\star} = \min_{\boldsymbol{X} = (\boldsymbol{X}_{2r}, \{\boldsymbol{X}_{g_j}\}_{j \in [l_g]})} \left\{ \langle \boldsymbol{C}_1, \boldsymbol{X}_{2r} \rangle \mid \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \boldsymbol{X} \succeq 0 \right\}, \tag{A8}$$

where the variable $\boldsymbol{X} = (\boldsymbol{X}_{2r}, \{\boldsymbol{X}_{g_j}\}_{j \in [l_g]})$ is a collection of positive-semidefinite matrices (*cf.* (A3) and (A9)), the objective is the one given in (A1), and the linear constraints $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$ collect all the constraints in (A4), (A5), and (A7). Problem (A8) can be readily formulated as a multi-block SDP in the primal form (SDP), which matches the data format used by common SDP solvers. The matrix $\boldsymbol{X}_{2r}$ solving (A8) is typically referred to as the *pseudo-moment matrix*.[36] One can solve the relaxation for different choices of r, leading to a *hierarchy* of convex relaxations.

While we presented Lasserre's hierarchy in a somewhat procedural way, the importance of the hierarchy lies in its stunning theoretical properties, that we review below.

Theorem A1 (Lasserre's Hierarchy [25, 104, 109]). *Let $-\infty < p^{\star} < \infty$ be the optimum of* (POP) *and $f_{2r}^{\star}$ (resp. $\boldsymbol{X}_{2r}^{\star}$) be the optimum (resp. one optimizer) of* (A8), *and assume* (POP) *is explicitly bounded (*i.e., *it satisfies the Archimedeanness condition in [48, Definition 3.137]), then*

(i) *(lower bound and convergence) $f_{2r}^{\star}$ converges to $p^{\star}$ from below as $r \to \infty$, and convergence occurs at a finite r under suitable technical conditions [109];*

---

[36] The rationale behind this name will become apparent in Appendix B.

(ii) *(rank-one solutions) if $f_{2r}^\star = p^\star$ at some finite r, then for every global minimizer $\boldsymbol{x}^\star$ of (POP),*
$\boldsymbol{X}_{2r}^\star \triangleq [\boldsymbol{x}^\star]_r[\boldsymbol{x}^\star]_r^\mathsf{T}$ is optimal for (A8), and every rank-one optimal solution $\boldsymbol{X}_{2r}^\star$ of (A8) can
be written as $[\boldsymbol{x}^\star]_r[\boldsymbol{x}^\star]_r^\mathsf{T}$ for some $\boldsymbol{x}^\star$ that is optimal for (POP);

(iii) *(optimality certificate) if* $\operatorname{rank}(\boldsymbol{X}_{2r}^\star) = 1$ *at some finite r, then $f_{2r}^\star = p^\star$.*

Theorem A1 states that (A8) provides a hierarchy of lower bounds for (POP). When the relaxation is exact ($p^\star = f_{2r}^\star$), global minimizers of (POP) correspond to rank-one solutions of (A8). Moreover, after solving the convex SDP (A8), one can check the rank of the optimal solution $\boldsymbol{X}_{2r}^\star$ to obtain a *certificate* of global optimality.

Further tightening the relaxation. As we discussed above, in the standard presentation of Lasserre's hierarchy, one adds a localizing matrix for each inequality constraint to enforce constraints such as $g_j \cdot \boldsymbol{X}_{2(r-\lceil \deg(g_j)/2\rceil)} \succeq 0$. However, in principle, we could also add constraints enforcing $g_{j_1} \cdot g_{j_2} \cdot \boldsymbol{M} \succeq 0$, for any pair of inequality constraints $g_{j_1} \geq 0$ and $g_{j_2} \geq 0$, for $j_1, j_2 \in [l_g]$. More generally, we can add constraints $\prod_{j \in \mathcal{S}} g_j \cdot \boldsymbol{M} \succeq 0$, for any subset $\mathcal{S} \subseteq [l_g]$ as long as $\deg\left(\prod_{j \in \mathcal{S}} g_j\right)$ has degree no larger than $2r$. After adding those extra constraints, we can still phrase the resulting relaxation in the standard form (SDP), by adding extra matrix variables $\boldsymbol{X}_{\mathcal{S}} = \prod_{j \in \mathcal{S}} g_j \cdot \boldsymbol{X}_{2(r-\lceil \sum_{j \in \mathcal{S}} \deg(g_j)/2\rceil)}$, and then forcing these matrices to be a linear function of $\boldsymbol{X}_{2r}$:

$$\text{localizing matrices}: \quad \boldsymbol{X}_{\mathcal{S}} \succeq 0, \quad \mathcal{S} \subseteq [l_g], \tag{A9}$$

$$\begin{aligned}\text{localizing constraints}: \langle \boldsymbol{A}_{\text{loc},\mathcal{S},kh}, \boldsymbol{X}_{2r}\rangle = [\boldsymbol{X}_{\mathcal{S}}]_{hk}, \\ \mathcal{S} \subseteq [l_g], \quad 1 \leq h \leq k \leq \underline{d}_{r-\lceil \sum_{j \in \mathcal{S}} \deg(g_j)/2\rceil},\end{aligned} \tag{A10}$$

where, similarly to the standard Lasserre's relaxation, the linear constraints (for some suitable matrices $\boldsymbol{A}_{\text{loc},\mathcal{S},kh}$) enforce each entry of $\boldsymbol{X}_{\mathcal{S}}$ to be a linear combination of the entries in $\boldsymbol{X}_{2r}$.

The additional constraints in eq. (A10) make the relaxation tighter compared to the standard presentation of Lasserre's relaxation, but are not necessary to obtain the convergence result in Theorem A1, which holds regardless for explicitly bounded constraint sets. However, these constraints become necessary to obtain convergence results akin to Theorem A1 for the case where the set of constraints is *not* explicitly bounded (see [50, Section 3.3] and [48, p. 115] for a more extensive discussion). For this reason, in order to maintain generality, related work following the "proofs to algorithms" paradigm typically assumes those constraints to be present, see, *e.g.,* [15, 37]. These terms will indeed appear in the definitions of sos proofs and constrained pseudo-distribution, see Appendix B. In order to keep the definitions in our monograph consistent with [15, 37], we will also assume these terms to be present, even though they are not strictly necessary under Assumption 2.

B   Pseudo-distributions and Moment Relaxations

The start of this section follows standard introductions about pseudo-distributions given in related work [15, 49, 78, 37], while later in the section we attempt to draw more explicit connections with the optimization machinery in Appendix A. Although such a connection is self-evident to the expert reader (indeed pseudo-distributions are the language traditionally used to justify the moment relaxation [104]), such a connection is often less immediate for the practitioner, in particular when taking the algorithmic view of moment relaxations presented in Appendix A.

Pseudo-distributions. Pseudo-distributions are a generalization of the concept of probability distribution. A standard probability distribution μ with finite support in $\mathbb{R}^{d_x}$ is simply a function

$\mu : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ such that $\sum_{\boldsymbol{x} \in \text{support}(\mu)} \mu(\boldsymbol{x}) = 1$ and $\mu(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$. In other words, if $\text{support}(\mu)$ is a finite collection a points in $\mathbb{R}^{d_x}$, μ assigns a non-negative probability mass to each of these points, such that those probabilities sum up to 1. Similarly, a pseudo-distribution $\tilde{\mu}$ is a finitely supported function such that $\sum_{\boldsymbol{x} \in \text{support}(\tilde{\mu})} \tilde{\mu}(\boldsymbol{x}) = 1$ but in this case the non-negativity condition is replaced by a milder condition (*i.e.*, a pseudo-distribution can assume negative values over its support).

In order to formally introduce the notion of pseudo-distribution, we start by defining the *pseudo-expectation* of a function $f : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ under a finitely supported function $\tilde{\mu}$:

$$\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f(\boldsymbol{x}) \right] \doteq \sum_{\boldsymbol{x} \in \text{support}(\tilde{\mu})} f(\boldsymbol{x}) \cdot \tilde{\mu}(\boldsymbol{x}). \tag{A11}$$

We are now ready to formally define a pseudo-distribution.

Definition A2 (Pseudo-distribution). *A finitely supported function $\tilde{\mu} : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ is a level-ℓ pseudo-distribution if $\tilde{\mathbb{E}}_{\tilde{\mu}} [1] = 1$ and $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f(\boldsymbol{x})^2 \right] \geq 0$ for all polynomials f of degree $\deg(f) \leq \ell/2$.*

In words, $\tilde{\mu}$ is a function that is allowed to become negative as long as its "expectation" (more precisely, pseudo-expectation) with respect to every squared polynomials $f(\boldsymbol{x})^2$ of sufficiently low degree remains positive. It is possible to show that a level-∞ pseudo-distribution is an actual probability distribution, since the condition $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f(\boldsymbol{x})^2 \right] \geq 0$ would enforce $\tilde{\mu}$ to remain positive (in this case the pseudo-expectation becomes the traditional expectation of the distribution).

Towards reconnecting pseudo-distributions with the optimization machinery in Appendix A, we start by observing the following link between pseudo-distributions and pseudo-moment matrices.

Lemma A3 (Pseudo-moment matrix [49]). *Let $\tilde{\mu} : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ be a finitely supported function with $\tilde{\mathbb{E}}_{\tilde{\mu}} [1] = 1$. Then, $\tilde{\mu}$ is a level-ℓ pseudo-distribution iff the pseudo-moment matrix $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[[\boldsymbol{x}]_{\ell/2} [\boldsymbol{x}]_{\ell/2}^{\mathsf{T}} \right]$ is positive semidefinite, where $[\boldsymbol{x}]_{\ell/2}$ is the vector of monomials of degree up to $\ell/2$.*

Now we define what it means for a pseudo-distribution to satisfy a set of polynomial constraints.

Definition A4 (Constrained pseudo-distribution). *Let $\mathcal{A} \doteq \{f_1 \geq 0, \ldots, f_m \geq 0\}$ be a set of polynomial constraints over $\mathbb{R}^{d_x}$. Let $\tilde{\mu} : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ be a level-ℓ pseudo-distribution. We say that $\tilde{\mu}$ satisfies $\mathcal{A}$ at degree k, denoted as $\tilde{\mu} \stackrel{|k}{\models} \mathcal{A}$, if every set $\mathcal{S} \subset [m]$ and every sum-of-squares polynomial h on $\mathbb{R}^{d_x}$ with $\deg(h) + \sum_{i \in \mathcal{S}} \max\{\deg(f_i), k\} \leq \ell$ satisfies:*

$$\tilde{\mathbb{E}}_{\tilde{\mu}} \left[h \cdot \prod_{i \in \mathcal{S}} f_i \right] \geq 0. \tag{A12}$$

Moreover, we say that $\tilde{\mu} \stackrel{|k}{\models} \mathcal{A}$ holds approximately *if the above inequalities are satisfied up to an error of $2^{-d_x^\ell} \cdot \|h\| \cdot \prod_{i \in \mathcal{S}} \|f_i\|$, where $\|\cdot\|$ denotes the Euclidean norm of the coefficients of the polynomial.*

The notion of pseudo-distributions *approximately* satisfying a set of constraints is useful to account for the practical observation that numerical SDP solvers (which we are going to use to find pseudo-distributions, as discussed later in this section) will only satisfy the constraints up to some numerical tolerance, and we have to make sure that such numerical errors do not lead us to draw incorrect conclusions using the sos proof system (see Appendix C).

In this monograph, we make use of the following facts about pseudo-distributions.

Fact A5 (Linearity [134]). *Let f, g be polynomials of degree at most ℓ in indeterminate $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and take $\alpha, \beta \in \mathbb{R}$. Then, for any level-$\ell$ pseudo-distribution $\tilde{\mu}$,*

$$\tilde{\mathbb{E}}_{\tilde{\mu}} \left[\alpha\, f(\boldsymbol{x}) + \beta\, g(\boldsymbol{x}) \right] = \alpha \tilde{\mathbb{E}}_{\tilde{\mu}} \left[f(\boldsymbol{x}) \right] + \beta \tilde{\mathbb{E}}_{\tilde{\mu}} \left[g(\boldsymbol{x}) \right]. \tag{A13}$$

Fact A6 (Cauchy-Schwarz for pseudo-distributions [37]). *Let f, g be polynomials of degree at most ℓ in indeterminate $\boldsymbol{x} \in \mathbb{R}^{d_x}$. Then, for any level-$\ell$ pseudo-distribution $\tilde{\mu}$,*

$$\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f \cdot g \right] \leq \sqrt{\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f^2 \right]} \cdot \sqrt{\tilde{\mathbb{E}}_{\tilde{\mu}} \left[g^2 \right]}, \tag{A14}$$

and (specializing the result above to $g = 1$):

$$\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f \right]^2 \leq \tilde{\mathbb{E}}_{\tilde{\mu}} \left[f^2 \right]. \tag{A15}$$

Fact A7 (Hölder's inequality for pseudo-distributions [15]). *Let f, g be sos polynomials. Let p, q be positive integers such that $1/p + 1/q = 1$. Then, for any pseudo-distribution $\tilde{\mu}$ of level $\ell \geq pq \cdot deg(f) \cdot deg(g)$, we have:*

$$\left(\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f \cdot g \right] \right)^{pq} \leq \tilde{\mathbb{E}}_{\tilde{\mu}} \left[f^p \right]^q \cdot \tilde{\mathbb{E}}_{\tilde{\mu}} \left[g^q \right]^p. \tag{A16}$$

In particular, for all even integers $k \geq 2$, and polynomial f with $deg(f) \cdot k \leq \ell$:

$$\left(\tilde{\mathbb{E}}_{\tilde{\mu}} \left[f \right] \right)^k \leq \tilde{\mathbb{E}}_{\tilde{\mu}} \left[f^k \right]. \tag{A17}$$

Fact A8 (Norm inequality for pseudo-distributions). *Let $\boldsymbol{v}$ be an m-vector with polynomial entries of degree at most $\ell/2$ in indeterminate $\boldsymbol{x} \in \mathbb{R}^{d_x}$. Then, for any degree-$\ell$ pseudo-distribution $\tilde{\mu}$,*

$$\left\| \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\boldsymbol{v} \right] \right\|_2^2 \leq \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\| \boldsymbol{v} \|_2^2 \right]. \tag{A18}$$

Proof. By definition, $\| \boldsymbol{v} \|_2^2 = \sum_{i=1}^m v_i^2$. Moreover, by (A17), $\left(\tilde{\mathbb{E}}_{\tilde{\mu}} \left[v_i \right] \right)^2 \leq \tilde{\mathbb{E}}_{\tilde{\mu}} \left[v_i^2 \right]$. Therefore:

$$\left\| \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\boldsymbol{v} \right] \right\|_2^2 = \sum_{i=1}^m \left(\tilde{\mathbb{E}}_{\tilde{\mu}} \left[v_i \right] \right)^2 \leq \sum_{i=1}^m \tilde{\mathbb{E}}_{\tilde{\mu}} \left[v_i^2 \right] \overset{linearity}{=\!=} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\sum_{i=1}^m v_i^2 \right] = \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\| \boldsymbol{v} \|_2^2 \right], \tag{A19}$$

proving the claim. ■

Making the connection with moment relaxations explicit. The non-expert reader might still be confused about the relation between pseudo-distributions and moment relaxations. To shed some light, let us restate our (POP):

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^{d_x}} \quad & p(\boldsymbol{x}) \\ \text{subject to} \quad & h_i(\boldsymbol{x}) = 0, \quad i = 1, \dots, l_h \\ & g_j(\boldsymbol{x}) \geq 0, \quad j = 1, \dots, l_g. \end{aligned} \tag{A20}$$

Now, we start by relaxing (A20) using pseudo-distributions, and show that this leads back to the relaxation presented in Appendix A. In particular, we relax (A20) to:

$$\min_{\tilde{\mu}} \ \tilde{\mathbb{E}}_{\tilde{\mu}}\left[p(\boldsymbol{x})\right] \tag{A21}$$

$$\text{subject to} \quad \tilde{\mu} \text{ is a level-}\ell \text{ pseudo-distribution} \tag{A22}$$

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[h_i(\boldsymbol{x}) \cdot q(\boldsymbol{x})\right] = 0, \tag{A23}$$

$$\text{for all } i = 1,\dots,l_h \text{ and for all } q \in \mathbb{R}[\boldsymbol{x}], \text{ such that } \deg\left(h_i \cdot q\right) \leq \ell$$

$$\tilde{\mathbb{E}}_{\mu}\left[\prod_{j \in \mathcal{S}} g_j(\boldsymbol{x}) \cdot s(\boldsymbol{x})^2\right] \geq 0, \tag{A24}$$

$$\text{for all } \mathcal{S} \subseteq [l_g] \text{ and for all } s \in \mathbb{R}[\boldsymbol{x}] \text{ such that } \deg\left(\prod_{j \in \mathcal{S}} g_j \cdot s^2\right) \leq \ell.$$

Despite the complexity of (A21), it is apparent that (A21) is a relaxation of (A20): for any $\boldsymbol{x}$ that is feasible for (A20) (*i.e.*, that satisfies $h_i(\boldsymbol{x}) = 0$ and $g_j(\boldsymbol{x}) \geq 0$), we can define a (pseudo-)distribution μ_x supported on $\boldsymbol{x}$ (*i.e.*, $\mu_x(\boldsymbol{x}) = 1$ and zero elsewhere) which is also feasible for (A21) (such pseudo-distribution μ_x is such that $\tilde{\mathbb{E}}_{\mu_x}\left[p(\boldsymbol{x})\right] = p(\boldsymbol{x})$ for any polynomial p, hence also preserving the objective of (A20)). Indeed, it is possible to show that if we require $\tilde{\mu}$ to be an actual distribution, and replace the pseudo-expectations with actual expectations, then (A21) becomes equivalent to (A20), see [108] for a more extensive discussion. The advantage of the relaxation (A21) is its tractability: while (A20) is NP-hard [108], the relaxation (A21) can be written as a semidefinite program (SDP) and solved in polynomial time. Indeed, in the following we show that rewriting (A21) as an SDP leads us back to the same moment relaxation we procedurally introduced in Appendix A. Towards this goal, we will need some extra notation.

Preliminaries to connect problem (A21) *with Appendix A:* Recall that $[\boldsymbol{x}]_{\ell/2}$ is the vector of monomials of degree up to $\ell/2$ and therefore the moment matrix $\boldsymbol{X}_\ell \triangleq [\boldsymbol{x}]_{\ell/2}[\boldsymbol{x}]_{\ell/2}^\mathsf{T}$ contains all monomials of degree up to ℓ. It will be useful to define (and visualize) the pseudo-expectation of the moment matrix: $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{X}_\ell\right]$. For instance, for the case with $\boldsymbol{x} = [x_1 ; x_2]$ and $\ell = 4$:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{X}_4\right] \triangleq \begin{bmatrix} 1 & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2^2\right] \\ \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^3\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2^2\right] \\ \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2^3\right] \\ \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^3\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^4\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^3 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2^2\right] \\ \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^3 x_2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2^3\right] \\ \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2^3\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2^2\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1 x_2^3\right] & \tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_2^4\right] \end{bmatrix}. \tag{A25}$$

In the following, we will also need a more convenient way to index the monomials in $[\boldsymbol{x}]_\ell$ (and, as a consequence, the entries of $\boldsymbol{X}_\ell$ and $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{X}_\ell\right]$). Using standard notation, for a vector $\boldsymbol{\alpha} \in \mathbb{N}^{d_x}$, we write $\boldsymbol{x}^{\boldsymbol{\alpha}}$ to denote the monomial with exponents $\boldsymbol{\alpha}$ (for instance, for $\boldsymbol{\alpha} = [1 ; 3 ; 0 ; 5]$, $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1 x_2^3 x_4^5$). We also denote $|\boldsymbol{\alpha}| \triangleq \sum_{i=1}^{d_x} \alpha_i$, which is the degree of the monomial. Using this notation, we can index with $\boldsymbol{\alpha}$ the monomials appearing in $[\boldsymbol{x}]_\ell$. For instance, for $\ell = 2$:

$$[\boldsymbol{x}]_2 \triangleq \big[\ \overbrace{1}^{[0 ; 0]}\ ;\ \overbrace{x_1}^{[1 ; 0]}\ ;\ \overbrace{x_2}^{[0 ; 1]}\ ;\ \overbrace{x_1^2}^{[2 ; 0]}\ ;\ \overbrace{x_1 x_2}^{[1 ; 1]}\ ;\ \overbrace{x_2^2}^{[0 ; 2]}\ \big], \tag{A26}$$

where for each monomial, we reported the corresponding "index" $\boldsymbol{\alpha}$. We can similarly index the rows and columns of $\boldsymbol{X}_\ell$ and $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{X}_\ell\right]$ using two indices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For instance, the entry of the matrix

indexed by row $\boldsymbol{\alpha} = [2\,;\,0]$ and column $\boldsymbol{\alpha} = [0\,;\,1]$ in (A25) will be $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[x_1^2 x_2\right]$. Note that the monomial appearing in row $\boldsymbol{\alpha}$ and column $\boldsymbol{\beta}$ of the moment matrix $\boldsymbol{X}_\ell$ will always have exponent $\boldsymbol{\alpha} + \boldsymbol{\beta}$, since, due to the definition of the moment matrix $(\boldsymbol{X}_\ell \triangleq [\boldsymbol{x}]_{\ell/2}[\boldsymbol{x}]_{\ell/2}^\mathsf{T})$ its entry $[\boldsymbol{X}_\ell]_{\boldsymbol{\alpha\beta}} = \boldsymbol{x}^{\boldsymbol{\alpha}} \cdot \boldsymbol{x}^{\boldsymbol{\beta}}$ and for two monomials $\boldsymbol{x}^{\boldsymbol{\alpha}}$ and $\boldsymbol{x}^{\boldsymbol{\beta}}$, it holds:[37]

$$\boldsymbol{x}^{\boldsymbol{\alpha}} \cdot \boldsymbol{x}^{\boldsymbol{\beta}} = \boldsymbol{x}^{\boldsymbol{\alpha}+\boldsymbol{\beta}}. \tag{A27}$$

Finally, we will conveniently use the following representation of a polynomial $f(\boldsymbol{x})$ of degree ℓ:

$$f(\boldsymbol{x}) = \textstyle\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\leq\ell} \bar{f}_{\boldsymbol{\alpha}}\, \boldsymbol{x}^{\boldsymbol{\alpha}}, \tag{A28}$$

where we simply observed that the polynomial is the sum of monomials $\boldsymbol{x}^{\boldsymbol{\alpha}}$ of degree $|\boldsymbol{\alpha}|\leq \ell$, and with suitable coefficients $\bar{f}_{\boldsymbol{\alpha}}$, again indexed by $\boldsymbol{\alpha}$.

We are now ready to show that both objective and constraints (A21) can be rewritten in a way that leads back to the moment relaxation in Appendix A.

Rewriting the objective (A21): To simplify the objective $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[p(\boldsymbol{x})\right]$, we note that the pseudo-expectation is a linear operator, hence:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[p(\boldsymbol{x})\right] \overset{\text{using (A28)}}{=} \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\leq\ell} \bar{p}_{\boldsymbol{\alpha}}\, \boldsymbol{x}^{\boldsymbol{\alpha}}\right] \overset{\text{using Fact A5}}{=} \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\leq\ell} \bar{p}_{\boldsymbol{\alpha}}\, \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{x}^{\boldsymbol{\alpha}}\right] \overset{\text{for a suitable matrix } \boldsymbol{C}_1}{=} \left\langle \boldsymbol{C}_1, \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{X}_\ell\right]\right\rangle, \tag{A29}$$

which indeed produces the same structure as the objective of the moment relaxation in (A1) with $\ell = 2r$; as we will see in a while, $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{X}_\ell\right]$ will become the main matrix variable in the optimization.

Rewriting the equality constraints (A23): To simplify the constraint $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[h_i(\boldsymbol{x})q(\boldsymbol{x})\right] = 0$ (which has to hold for polynomials q of degree $\deg(h_i \cdot q) \leq \ell$) we note that it suffices to require $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[h_i(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\beta}}\right] = 0$ for $|\boldsymbol{\beta}|\leq \ell - \deg(h_i)$; this follows from the fact that any polynomial is a sum of monomials and the pseudo-expectation is a linear function. Let us now manipulate $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[h_i(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\beta}}\right] = 0$ as follows:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[h_i(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\beta}}\right] = 0 \overset{\text{using (A28) for } h_i(\boldsymbol{x})}{\Longleftrightarrow} \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{\boldsymbol{\alpha}} \bar{h}_{i,\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\beta}}\right] = 0 \overset{\text{using (A27)}}{\Longleftrightarrow} \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{\boldsymbol{\alpha}} \bar{h}_{i,\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}+\boldsymbol{\beta}}\right] = 0$$
$$\overset{\text{using Fact A5}}{\Longleftrightarrow} \sum_{\boldsymbol{\alpha}} \bar{h}_{i,\boldsymbol{\alpha}}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{x}^{\boldsymbol{\alpha}+\boldsymbol{\beta}}\right] = 0 \overset{\text{for a suitable matrix } \boldsymbol{A}_{i,\boldsymbol{\beta}}}{\Longleftrightarrow} \left\langle \boldsymbol{A}_{i,\boldsymbol{\beta}}, \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{X}_\ell\right]\right\rangle = 0, \tag{A30}$$

which has to be imposed for each $\boldsymbol{\beta}$ such that $|\boldsymbol{\beta}|\leq \ell - \deg(h_i)$. Note that the constraints in (A30) capture both the equality constraints in (A2) (for $|\boldsymbol{\beta}|= 0$) as well as the redundant constraints (A5) (for $0 < |\boldsymbol{\beta}|\leq \ell - \deg(h_i)$).

Rewriting the inequality constraints in (A46): We simplify the constraint $\tilde{\mathbb{E}}_{\mu}\left[\prod_{j\in\mathcal{S}} g_j(\boldsymbol{x}) \cdot s(\boldsymbol{x})^2\right] \geq 0$, which has to hold for all $\mathcal{S} \subseteq [l_g]$ and for all $s \in \mathbb{R}[\boldsymbol{x}]$ such that $\deg\left(\prod_{j\in\mathcal{S}} g_j \cdot s^2\right) \leq \ell$. Towards this goal, we use the representation (A28) for $s(\boldsymbol{x})$ and write $s(\boldsymbol{x}) = \sum_{\boldsymbol{\beta}:|\boldsymbol{\beta}|\leq t} \bar{s}_{\boldsymbol{\beta}}\boldsymbol{x}^{\boldsymbol{\beta}}$, where

---

[37]For instance, the product between the monomial $x_1 x_2^3 x_4^5$ (namely, $\boldsymbol{x}^{\boldsymbol{\alpha}}$ with $\boldsymbol{\alpha} = [1\,;\,3\,;\,0\,;\,5]$) and the monomial $x_1^2 x_2 x_3$ (namely, $\boldsymbol{x}^{\boldsymbol{\beta}}$ with $\boldsymbol{\beta} = [2\,;\,1\,;\,1\,;\,0]$) is $(x_1 x_2^3 x_4^5) \cdot (x_1^2 x_2 x_3) = x_1^3 x_2^4 x_3 x_4^5$, which corresponds to the exponent vector $[3\,;\,4\,;\,1\,;\,5]$.

$t \triangleq \left\lfloor \frac{\ell - \deg\left(\prod_{j \in \mathcal{S}} g_j\right)}{2} \right\rfloor$. Therefore, we obtain:

$$\tilde{\mathbb{E}}_\mu \left[g_j(\boldsymbol{x}) \cdot s(\boldsymbol{x})^2 \right] \geq 0 \overset{\text{expanding } s^2}{\Longleftrightarrow} \tilde{\mathbb{E}}_\mu \left[g_j(\boldsymbol{x}) \cdot \sum_{\boldsymbol{\alpha}: |\boldsymbol{\alpha}| \leq t} \bar{s}_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}: |\boldsymbol{\beta}| \leq t} \bar{s}_{\boldsymbol{\beta}} \boldsymbol{x}^{\boldsymbol{\beta}} \right] \geq 0 \qquad \text{(A31)}$$

$$\overset{\text{rearranging}}{\Longleftrightarrow} \tilde{\mathbb{E}}_\mu \left[\sum_{\boldsymbol{\alpha}, \boldsymbol{\beta}: |\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq t} \bar{s}_{\boldsymbol{\alpha}} \bar{s}_{\boldsymbol{\beta}} \boldsymbol{x}^{\boldsymbol{\alpha} + \boldsymbol{\beta}} g_j(\boldsymbol{x}) \right] \geq 0 \qquad \text{(A32)}$$

$$\overset{\text{using (A28) on } \prod_{\mathcal{S} \subseteq [l_g]} g_j(\boldsymbol{x})}{\Longleftrightarrow} \tilde{\mathbb{E}}_\mu \left[\sum_{\boldsymbol{\alpha}, \boldsymbol{\beta}: |\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq t} \bar{s}_{\boldsymbol{\alpha}} \bar{s}_{\boldsymbol{\beta}} \boldsymbol{x}^{\boldsymbol{\alpha} + \boldsymbol{\beta}} \sum_{\boldsymbol{\gamma}: |\boldsymbol{\gamma}| \leq \deg\left(\prod_{j \in \mathcal{S}} g_j(\boldsymbol{x})\right)} \bar{g}_{\mathcal{S}, \boldsymbol{\gamma}} \boldsymbol{x}^{\boldsymbol{\gamma}} \right] \geq 0 \qquad \text{(A33)}$$

$$\overset{\text{rearranging}}{\Longleftrightarrow} \tilde{\mathbb{E}}_\mu \left[\sum_{\boldsymbol{\alpha}, \boldsymbol{\beta}: |\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq t} \bar{s}_{\boldsymbol{\alpha}} \bar{s}_{\boldsymbol{\beta}} \sum_{\boldsymbol{\gamma}: |\boldsymbol{\gamma}| \leq \deg\left(\prod_{j \in \mathcal{S}} g_j(\boldsymbol{x})\right)} \bar{g}_{\mathcal{S}, \boldsymbol{\gamma}} \boldsymbol{x}^{\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma}} \right] \geq 0 \qquad \text{(A34)}$$

$$\overset{\text{using Fact A5}}{\Longleftrightarrow} \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta}: |\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq t} \bar{s}_{\boldsymbol{\alpha}} \bar{s}_{\boldsymbol{\beta}} \sum_{\boldsymbol{\gamma}: |\boldsymbol{\gamma}| \leq \deg\left(\prod_{j \in \mathcal{S}} g_j(\boldsymbol{x})\right)} \bar{g}_{\mathcal{S}, \boldsymbol{\gamma}} \tilde{\mathbb{E}}_\mu \left[\boldsymbol{x}^{\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma}} \right] \geq 0. \qquad \text{(A35)}$$

Now note that $|\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma}| \leq \ell$ by construction, and hence we can write, for a given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, each $\sum_{\boldsymbol{\gamma}: |\boldsymbol{\gamma}| \leq \deg\left(\prod_{j \in \mathcal{S}} g_j(\boldsymbol{x})\right)} \bar{g}_{\mathcal{S}, \boldsymbol{\gamma}} \tilde{\mathbb{E}}_\mu \left[\boldsymbol{x}^{\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma}} \right]$ as a linear function of $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[\boldsymbol{X}_\ell \right]$. Moreover, since $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are such that $|\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq t$, we can group these entries into an $t \times t$ matrix $\boldsymbol{X}_{\mathcal{S}}$, which is such that:

$$[\boldsymbol{X}_{\mathcal{S}}]_{\boldsymbol{\alpha}, \boldsymbol{\beta}} = \left\langle \boldsymbol{A}_{\mathrm{loc}, \mathcal{S}, \boldsymbol{\alpha}\boldsymbol{\beta}}, \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\boldsymbol{X}_\ell \right] \right\rangle, \qquad \text{(A36)}$$

for some suitable matrix $\boldsymbol{A}_{\mathrm{loc}, \mathcal{S}, \boldsymbol{\alpha}\boldsymbol{\beta}}$, such that $\left\langle \boldsymbol{A}_{\mathrm{loc}, \mathcal{S}, \boldsymbol{\alpha}\boldsymbol{\beta}}, \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\boldsymbol{X}_\ell \right] \right\rangle = \sum_{\boldsymbol{\gamma}: |\boldsymbol{\gamma}| \leq \deg\left(\prod_{j \in \mathcal{S}} g_j\right)} \bar{g}_{i, \boldsymbol{\gamma}} \tilde{\mathbb{E}}_\mu \left[\boldsymbol{x}^{\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma}} \right]$. Using the matrix $\boldsymbol{X}_{\mathcal{S}}$ and defining a vector $\bar{\boldsymbol{s}} \in \mathbb{R}^t$ with entries $\bar{s}_{\boldsymbol{\alpha}}$ for $|\boldsymbol{\alpha}| \leq t$, we rewrite (A35) as:

$$\sum_{\boldsymbol{\alpha}, \boldsymbol{\beta}: |\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq t} \bar{s}_{\boldsymbol{\alpha}} \bar{s}_{\boldsymbol{\beta}} [\boldsymbol{X}_{\mathcal{S}}]_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \geq 0 \iff \bar{\boldsymbol{s}}^{\mathsf{T}} \boldsymbol{X}_{\mathcal{S}} \bar{\boldsymbol{s}} \geq 0. \qquad \text{(A37)}$$

Since this has to hold for any $\bar{\boldsymbol{s}}$ (*i.e.*, any polynomial $s(\boldsymbol{x})$ of appropriate degree), we conclude the constraint above is equivalent to:

$$\boldsymbol{X}_{\mathcal{S}} \succeq 0. \qquad \text{(A38)}$$

Now we can easily see that (A36) and (A38) match the localizing constraints we wrote in (A7).

Rewriting (A22): Finally, the constraint (A22) imposes that $\tilde{\mu}$ must be a level ℓ pseudo-distribution. However, we know from Lemma A3 that $\tilde{\mu}$ is a level-ℓ pseudo-distribution if and only if the pseudo-moment matrix $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[[\boldsymbol{x}]_{\ell/2} [\boldsymbol{x}]_{\ell/2}^{\mathsf{T}} \right]$ is positive semidefinite and $\tilde{\mathbb{E}}_{\tilde{\mu}} [1] = 1$. Therefore, we can reparametrize the objective (A29) and constraints (A30), (A36), (A38) with a matrix variable (in place of $\tilde{\mathbb{E}}_{\tilde{\mu}} [\boldsymbol{X}_\ell]$) that is constrained to be positive semidefinite and to have the top-left entry equal to 1 (*cf.* (A25)). This yields back the relaxation described in Appendix A as expected.

Constrained pseudo-distributions: a practical view. So far we have shown that taking suitable pseudo-expectations over the objective and constraints in a polynomial optimization problem leads to a convex relaxation, known as the moment relaxation. Now we want to shed some light on Definition A4 by showing that the condition (A12) is indeed the same as the inequality constrains in (A46) and hence admits the same transcription as an SDP.

Towards this goal, let us consider the following *feasibility* POP:

$$\text{find} \quad \boldsymbol{x} \in \mathbb{R}^{d_x} \tag{A39}$$

$$\text{subject to} \quad h_i(\boldsymbol{x}) = 0, i = 1, \ldots, l_h \tag{A40}$$

$$g_j(\boldsymbol{x}) \geq 0, j = 1, \ldots, l_g. \tag{A41}$$

This is similar to (POP), with the exception that we are looking for a feasible solution rather than optimizing a cost function. Now note that we can write a polynomial equality $h_i(\boldsymbol{x}) = 0$ as two inequality constraints $h_i(\boldsymbol{x}) \leq 0$ and $-h_i(\boldsymbol{x}) \leq 0$. Hence, without loss of generality we rewrite (A39) as:

$$\text{find} \quad \boldsymbol{x} \in \mathbb{R}^{d_x} \tag{A42}$$

$$\text{subject to} \quad f_j(\boldsymbol{x}) \geq 0, \quad j = 1, \ldots, m, \tag{A43}$$

for suitable polynomials f_i, $i = 1, \ldots, m$. Similarly to what we did earlier in this section, we relax (A42) by using pseudo-expectations:

$$\text{find} \quad \tilde{\mu} \tag{A44}$$

$$\text{subject to} \quad \tilde{\mu} \text{ is a level-}\ell \text{ pseudo-distribution} \tag{A45}$$

$$\tilde{\mathbb{E}}_\mu \left[s(\boldsymbol{x})^2 \cdot \prod_{i \in \mathcal{S}} f_i(\boldsymbol{x}) \right] \geq 0, \tag{A46}$$

for every set $\mathcal{S} \in [m]$ and every $s \in \mathbb{R}[\boldsymbol{x}]$ such that $\deg\left(s^2 \cdot \prod_{i \in \mathcal{S}} f_i\right) \leq \ell$.

First of all, we note that (A46) matches the definition of constrained pseudo-distribution in Definition A4 for $k = 0$. Moreover, following the same derivation as above, we can easily show that (i) (A44) can be transcribed as a standard SDP, and (ii) every pseudo-distribution solving Lasserre's relaxation of a (POP) satisfies the set of constraints in the (POP) in the sense of Definition A4.

Now we note that Definition A4 allows some extra slack through the parameter k, *i.e.,* $\tilde{\mu}$ satisfies $\mathcal{A}$ at degree k, if every set $\mathcal{S} \subset [m]$ and every sum-of-squares polynomial h on $\mathbb{R}^{d_x}$ with $\deg\left(s^2\right) + \sum_{i \in \mathcal{S}} \max\{\deg\left(f_i\right), k\} \leq \ell$ satisfies $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[s^2 \cdot \prod_{i \in \mathcal{S}} f_i\right] \geq 0$. This essentially means that the inequality $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[s^2 \cdot \prod_{i \in \mathcal{S}} f_i\right] \geq 0$ is enforced for a smaller number of subsets $\mathcal{S}$.

C   Sum-of-Squares Proofs

Sum-of-squares proofs provide an advanced way to reason about polynomial constraints and to infer properties of pseudo-distributions, or, equivalently, properties of the moment relaxation in Appendix A. The presentation in this section builds on [49], but also collects inference rules from other papers, which we cite as we present the results.

Let us denote with $f(\boldsymbol{x})$ a polynomial in variables $\boldsymbol{x} = [x_1; x_2; \ldots; x_{d_x}]$ and let $\mathcal{A} = \{f_1(\boldsymbol{x}) \geq 0, \ldots, f_m(\boldsymbol{x}) \geq 0\}$ be a system of polynomial constraints over $\mathbb{R}^{d_x}$. In the following, we omit the argument when clear from the context and write f instead of $f(\boldsymbol{x})$. A polynomial p is sum-of-squares (sos) if there exist polynomials $q_1, \ldots, q_t$ such that $p = q_1^2 + \ldots + q_t^2$.

The key idea is to relate two sets of polynomial constraints using a "sum-of-squares proof" (the definition below is the same as Definition 3 in the main manuscript).

Definition A9 (Sum-of-squares proof). *Given a system of polynomial constraint $\mathcal{A}$ and a polynomial g, a sum-of-square (sos) proof that the system $\mathcal{A}$ implies $g \geq 0$ consists of sum-of-squares polynomials $\{p_{\mathcal{S}}\}_{\mathcal{S} \subseteq [m]}$ such that:*

$$g = \sum_{\mathcal{S} \subseteq [m]} p_{\mathcal{S}} \cdot \prod_{i \in \mathcal{S}} f_i. \tag{A47}$$

We say that the proof has degree k if for every $\mathcal{S} \subseteq [m]$, $\deg\left(p_{\mathcal{S}} \cdot \prod_{i \in \mathcal{S}} f_i\right) \leq k$ where $\deg(\cdot)$ denotes the degree of a polynomial. We use the notation:

$$\mathcal{A}(\boldsymbol{x}) \left|\frac{k}{\boldsymbol{x}}\right. \{g(\boldsymbol{x}) \geq 0\} \qquad or \qquad \{f_i(\boldsymbol{x}) \geq 0, \ldots, f_m(\boldsymbol{x}) \geq 0\} \left|\frac{k}{\boldsymbol{x}}\right. \{g(\boldsymbol{x}) \geq 0\} \tag{A48}$$

to denote that there is a proof of degree at most k of the fact that $\mathcal{A} = \{f_i(\boldsymbol{x}) \geq 0, \ldots, f_m(\boldsymbol{x}) \geq 0\}$ implies $g \geq 0$ (i.e., any $\boldsymbol{x}$ that satisfies $\mathcal{A}(\boldsymbol{x})$ is such that $g(\boldsymbol{x}) \geq 0$). We omit the variables and write $\mathcal{A}(\boldsymbol{x}) \left|\frac{k}{}\right. \{g(\boldsymbol{x}) \geq 0\}$, when they are clear from the context. Moreover, we write

$$\left|\frac{k}{\boldsymbol{x}}\right. \{g(\boldsymbol{x}) \geq 0\} \tag{A49}$$

if there is a sum-of-squares proof that $g(\boldsymbol{x}) \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^{d_x}$ (i.e., $g(\boldsymbol{x})$ is sum-of-squares).

From eq. (A47), it is clear why the polynomials $p_{\mathcal{S}}$ are a "proof" of $g \geq 0$ for any $\boldsymbol{x}$ satisfying $\mathcal{A}$: for any $\boldsymbol{x} \in \mathcal{A}$, $\prod_{i \in \mathcal{S}} f_i \geq 0$ by definition, hence if we can write g as the product of a sum-of-squares (hence non-negative) polynomial and $\prod_{i \in \mathcal{S}} f_i$, we automatically prove that $g \geq 0$ whenever $\boldsymbol{x} \in \mathcal{A}$.

Sum-of-squares proofs allow us to deduce properties of pseudo-distributions: in particular, if we have an sos proof relating two sets of constraints, we can conclude that any pseudo-distribution satisfying a set of constraints, must also satisfy the other. This is formalized below.

Fact A10 (Soundness [37]). *Consider a level-ℓ pseudo-distribution $\tilde{\mu}$ such that $\tilde{\mu} \left|\frac{k}{}\right.\!\!= \mathcal{A}$. If there exists a sum-of-squares proof that $\mathcal{A} \left|\frac{k'}{}\right. \mathcal{B}$, then $\tilde{\mu} \left|\frac{k \cdot k' + k'}{}\right.\!\!= \mathcal{B}$.*

If the pseudo-distribution $\tilde{\mu}$ satisfies $\mathcal{A}$ only approximately, soundness continues to hold but we require an upper bound on the bit-complexity of the sum-of-squares proof $\mathcal{A} \left|\frac{k'}{}\right. \mathcal{B}$ (*i.e.*, the number of bits to write down the proof). In this monograph, we mostly disregard bit-complexity issues and refer the reader to [50, §3] for a more formal discussion. In other words, similarly to [15, 37], we assume that all numbers appearing in the input have bit complexity $d_x^{O(1)}$ and all sos proofs will have bit complexity $d_x^{O(\ell)}$, which is enough to claim soundness for the sos proof system.

Not only sos proofs allow us to infer properties of pseudo-distributions, but also the reverse is true. The following fact states that every property of a pseudo-distribution can be derived via a sum-of-squares proof.

Fact A11 (Completeness [37]). *Suppose $\ell \geq k' \geq k$ and $\mathcal{A}$ is a system of explicitly bounded polynomial constraints with degree at most k (i.e., $\mathcal{A} \left|-\right. \{\|\boldsymbol{x}\|_2^2 \leq M_x^2\}$ for some finite M_x). Let $\{g \geq 0\}$ be a polynomial constraint. If every level-ℓ pseudo-distribution that satisfies $\tilde{\mu} \left|\frac{k}{}\right.\!\!= \mathcal{A}$ also satisfies $\tilde{\mu} \left|\frac{k'}{}\right.\!\!= \mathcal{B}$, then for every $\epsilon > 0$ there is a sum-of-squares proof $\mathcal{A} \left|\frac{\ell}{}\right. \{g \geq -\epsilon\}$.*

Sos rules. Sum-of-squares provide a proof system to reasons about polynomial constraints. For instance, if we have an sos proof that $\mathcal{A}$ implies $g \geq 0$, we may want to use such a proof system

to infer if another implication also holds true, say $\mathcal{A}$ implies $g' \geq 0$ (for some other polynomial g'). Reasoning in this proof system is not immediate. For instance, the fact that $p(\boldsymbol{x}) \geq 0$ for some degree-k polynomial does not necessarily imply that there is a sum-of-squares proof $\vdash_{\boldsymbol{x}}^{k} \{p(\boldsymbol{x}) \geq 0\}$. Similarly, for some polynomial constraints $\mathcal{A}$ and polynomials $g(\boldsymbol{x})$ and $g(\boldsymbol{x})'$ with $g'(\boldsymbol{x}) \geq g(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^{d_x}$, the fact that $\mathcal{A} \vdash_{\boldsymbol{x}}^{k} \{g(\boldsymbol{x}) \geq 0\}$ does not necessarily imply that $\mathcal{A} \vdash_{\boldsymbol{x}}^{k} \{g'(\boldsymbol{x}) \geq 0\}$, since the latter fact might not admit a sum-of-squares proof. In this sense, the sos proof system is more restrictive than the typical algebraic manipulation we are used to. Fortunately, previous work provides a toolkit of inference rules that can be used to correctly reason over sos proofs. We collect key facts below, mostly drawing from [37, 15, 112, 113, 114].

Fact A12 (Inference Rules [37]). *The following inference rules hold for systems of polynomial constraints $\mathcal{A}, \mathcal{B}, \mathcal{C}$ and polynomials $f, g : \mathbb{R}^{d_x} \mapsto \mathbb{R}$:*

$$\text{addition:} \quad \frac{\mathcal{A} \vdash^{k} \{f \geq 0, g \geq 0\}}{\mathcal{A} \vdash^{k} \{f + g \geq 0\}} \tag{A50}$$

$$\text{multiplication:} \quad \frac{\mathcal{A} \vdash^{k} \{f \geq 0\}, \ \mathcal{A} \vdash^{k'} \{g \geq 0\}}{\mathcal{A} \vdash^{k+k'} \{f \cdot g \geq 0\}} \tag{A51}$$

$$\text{transitivity:} \quad \frac{\mathcal{A} \vdash^{k} \mathcal{B}, \ \mathcal{B} \vdash^{k'} \mathcal{C}}{\mathcal{A} \vdash^{k \cdot k'} \mathcal{C}} \tag{A52}$$

where, for two logical statements A *and* B*, we use the standard inference notation* $\frac{\mathsf{A}}{\mathsf{B}}$ *to denote that if* A *is true, then* B *must be true.*

Fact A13 (Basics, p. 59 in [48] and p. 70 in [50]). *Let $p(\boldsymbol{x})$ be a degree-k polynomial such that $p(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^{d_x}$. Then:*

$$\vdash_{\boldsymbol{x}}^{k} \{p(\boldsymbol{x}) \geq 0\} \tag{A53}$$

(i.e., $p(\boldsymbol{x})$ is sos) if:

- $d_x = 1$ *(univariate case),*
- $k = 2$ *(quadratic polynomials), or*
- $d_x = 2$ *and $k = 4$ (bivariate, quartic polynomials).*

Moreover, (A53) holds whenever p is a function over the Boolean hypercube $p : \{0, 1\}^{d_x} \mapsto \mathbb{R}$.

Fact A14 (Univariate polynomials over interval, Fact 3.7 in [37]). *For any univariate degree k polynomial $p(x) \geq 0$ for $x \in [a, b]$,*

$$\{x \geq a, x \leq b\} \vdash_{x}^{k} \{p(\boldsymbol{x}) \geq 0\}. \tag{A54}$$

Fact A15 (Sos generalized triangle inequality, Fact 4.8 in [15]). *For any $a_1, a_2, \ldots, a_m$*

$$\vdash_{a_1, a_2, \ldots, a_m}^{k} \left\{ \left(\sum_{i=1}^{m} a_i \right)^{k} \leq m^{k} \left(\sum_{i=1}^{m} a_i^{k} \right) \right\}. \tag{A55}$$

Fact A16 (Sos triangle inequality (same as Fact A15 with $m = 2$ and $k = 2$)). *For any a_1, a_2*

$$\left|\frac{2}{a,b}\right. \left\{ (a + b)^2 \leq 2^2 a^2 + 2^2 b^2 \right\}. \tag{A56}$$

Fact A17 (Sos triangle inequality 2.0, p. 18 in [15]). *For any indeterminates a, b, scalar δ, and even integer k:*

$$\left|\frac{k}{a,b}\right. \delta^k a^k \leq (2\delta)^k (a - b)^k + (2\delta)^k b^k. \tag{A57}$$

Fact A18 (Sos squaring). *Let f, g be sos polynomials of degree at most k and $\mathcal{A} = \{f_1(\boldsymbol{x}) \geq 0, \ldots, f_m(\boldsymbol{x}) \geq 0\}$ be a system of polynomial inequalities. If $\mathcal{A} \left|\frac{k'}{\boldsymbol{x}}\right. \{f \geq g\}$, then $\mathcal{A} \left|\frac{k'+k}{\boldsymbol{x}}\right. \{f^2 \geq g^2\}$.*

Proof. The assumption $\mathcal{A} \left|\frac{k}{\boldsymbol{x}}\right. f = 0$ implies that:

$$f - g = \sum_{\mathcal{S} \subseteq [m]} p_{\mathcal{S}} \cdot \prod_{i \in \mathcal{S}} f_i. \tag{A58}$$

Now note that $f^2 - g^2 = (f - g)(f + g)$ hence:

$$f^2 - g^2 = (f - g)(f + g) = (f + g) \sum_{\mathcal{S} \subseteq [m]} p_{\mathcal{S}} \cdot \prod_{i \in \mathcal{S}} f_i. \tag{A59}$$

Since f, g are sos polynomial, the previous relation proves $\mathcal{A} \left|\frac{k'+k}{\boldsymbol{x}}\right. \{f^2 \geq g^2\}$ with sos proof $(f + g) \cdot p_{\mathcal{S}}$ and by noting that the maximum degree appearing in (A59) is $k' + k$. ∎

Fact A19 (Sos triangle inequality with norms, Fact A.2 in [112]). *Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be n-length vectors of indeterminates. Then:*

$$\left|\frac{2}{\boldsymbol{x}_1, \boldsymbol{x}_2}\right. \left\{ \|\boldsymbol{x}_1 + \boldsymbol{x}_2\|_2^2 \leq 2 \|\boldsymbol{x}_1\|_2^2 + 2 \|\boldsymbol{x}_2\|_2^2 \right\}. \tag{A60}$$

Fact A20 (Sos generalized triangle inequality with norms). *Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be n-length vectors of indeterminates and $k \in \mathbb{N}$ be even. Then:*

$$\left|\frac{k}{\boldsymbol{x}_1, \boldsymbol{x}_2}\right. \left\{ \|\boldsymbol{x}_1 + \boldsymbol{x}_2\|_2^k \leq 2^k \|\boldsymbol{x}_1\|_2^k + 2^k \|\boldsymbol{x}_2\|_2^k \right\}. \tag{A61}$$

Proof.

$$\left|\frac{k}{\boldsymbol{x}_1, \boldsymbol{x}_2}\right. \|\boldsymbol{x}_1 + \boldsymbol{x}_2\|_2^k = \left(\|\boldsymbol{x}_1 + \boldsymbol{x}_2\|_2^2 \right)^{\frac{k}{2}} \overset{using\ (A60)}{\leq} \left(2 \|\boldsymbol{x}_1\|_2^2 + 2 \|\boldsymbol{x}_2\|_2^2 \right)^{\frac{k}{2}} \tag{A62}$$

$$\overset{using\ (A55)}{\leq} 2^{\frac{k}{2}} (2 \|\boldsymbol{x}_1\|_2^2)^{\frac{k}{2}} + 2^{\frac{k}{2}} (2 \|\boldsymbol{x}_2\|_2)^{\frac{k}{2}} = 2^k \|\boldsymbol{x}_1\|_2^k + 2^k \|\boldsymbol{x}_2\|_2^k. \tag{A63}$$

∎

Fact A21 (Sos Cauchy-Schwarz, Fact A.1 in [112]). *Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be polynomials in some indeterminates. Then:*

$$\left|\frac{4}{x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n}\right. \left\{ \left(\sum_{i=1}^{n} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{i=1}^{n} y_i^2 \right) \right\}, \tag{A64}$$

or, written in vector form, for two n-length vectors $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\vdash_{\boldsymbol{x},\boldsymbol{y}}^{4} \left\{ \left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}\right)^2 \leq \left(\|\boldsymbol{x}\|_2^2\right)\left(\|\boldsymbol{y}\|_2^2\right) \right\}. \tag{A65}$$

Fact A22 (Sos Hölder's inequality, Fact 4.4 in [15]). *Let f_i, g_i for $1 \leq i \leq n$ be sos polynomials. Let p, q be integers such that $\frac{1}{p} + \frac{1}{q} = 1$. Then:*

$$\vdash^{pq} \left\{ \left(\frac{1}{n}\sum_{i=1}^{n} f_i g_i\right)^{pq} \leq \left(\frac{1}{n}\sum_{i=1}^{n} f_i^p\right)^q \left(\frac{1}{n}\sum_{i=1}^{n} g_i^q\right)^p \right\}. \tag{A66}$$

Fact A23 (Sos Hölder's inequality 2.0, Fact A.6 in [112]). *Let $\omega_1, \ldots, \omega_n$ and $x_1, \ldots, x_n$ be indeterminates. Let $q \in \mathbb{N}$ be a power of 2. Then:*

$$\left\{\omega_i^2 = \omega_i, \forall i \in [n]\right\} \vdash_{\omega_1,\ldots,\omega_n,x_1,\ldots,x_n}^{O(q)} \left\{ \left(\sum_{i=1}^{n} \omega_i x_i\right)^q \leq \left(\sum_{i=1}^{n} \omega_i^2\right)^{q-1} \left(\sum_{i=1}^{n} x_i^q\right) \right\}, \tag{A67}$$

and

$$\left\{\omega_i^2 = \omega_i, \forall i \in [n]\right\} \vdash_{\omega_1,\ldots,\omega_n,x_1,\ldots,x_n}^{O(q)} \left\{ \left(\sum_{i=1}^{n} \omega_i x_i\right)^q \leq \left(\sum_{i=1}^{n} \omega_i^2\right)^{q-1} \left(\sum_{i=1}^{n} \omega_i x_i^q\right) \right\}. \tag{A68}$$

Fact A24 (Sos Hölder's inequality 3.0, Fact A.3 in [114]). *Let f_i, g_i for $1 \leq i \leq n$ be indeterminates. Then:*

$$\vdash^2 \left\{ \left(\frac{1}{n}\sum_{i=1}^{n} f_i g_i\right)^2 \leq \left(\frac{1}{n}\sum_{i=1}^{n} f_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} g_i^2\right) \right\}. \tag{A69}$$

Fact A25 (Lemma A.3 in [113]). *Let x be indeterminate and a be a positive real number. Then:*

$$\{x^2 \leq a^2\} \vdash \{x \leq a, x \geq -a\}. \tag{A70}$$

Fact A26 (Lemma A.2 in [113]). *Let $\boldsymbol{x}$ be indeterminate and $\boldsymbol{a}$ be a unit vector. Let $\mathcal{A}\{\|\boldsymbol{x}\|^2 = 1, (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{a})^2 \leq \tau\}$. Then, for any $\boldsymbol{b}$ such that $\|\boldsymbol{a} - \boldsymbol{b}\|_2 \leq 2\delta$, we have:*

$$\mathcal{A} \vdash \left\{ (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{b})^2 \leq (\sqrt{\tau} + \sqrt{\delta})^2 \right\}. \tag{A71}$$

We conclude with a self-evident fact that reassures us that certain manipulations of polynomials are easy to reason over, even in the sos proof system.

Fact A27 (Equalities). *Let f, g be polynomials and $\mathcal{A}$ be a system of polynomial inequalities. If $f = g$ and $\mathcal{A} \vdash_{\boldsymbol{x}}^{k} f = 0$, then $\mathcal{A} \vdash_{\boldsymbol{x}}^{k} g = 0$.*

D   Proof of Proposition 6: A Posteriori Contract for (TLS)

We start by restating the theorem for the reader's convenience.

> **Proposition A28** (Restatement of Proposition 6). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and denote with γ° the squared residual error of the ground truth $\boldsymbol{x}^\circ$ over the set of inliers $\mathcal{I}$, i.e., $\gamma^\circ \triangleq \sum_{i \in \mathcal{I}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ \right\|_2^2$. Moreover, assume the measurement set contains at least $\frac{n+\bar{d}}{2} + \frac{\gamma^\circ}{\bar{c}^2}$ inliers, where $\bar{d}$ is the size of a minimal set, and that every subset of $\bar{d}$ inliers is nondegenerate. Then, for any integer $d_{\mathcal{J}}$ such that $\bar{d} \le d_{\mathcal{J}} \le (2\alpha - 1)n - \frac{\gamma^\circ}{\bar{c}^2}$, an optimal solution $\boldsymbol{x}_{\mathsf{TLS}}$ of (TLS) satisfies*
>
> $$\left\| \boldsymbol{x}_{\mathsf{TLS}} - \boldsymbol{x}^\circ \right\|_2 \le \frac{2\sqrt{d_{\mathcal{J}}}\, \bar{c}}{\min_{\mathcal{J} \subset \mathcal{I}_{\mathsf{TLS}}, |\mathcal{J}| = d_{\mathcal{J}}} \sigma_{\min}(\boldsymbol{A}_{\mathcal{J}})}, \tag{A72}$$
>
> *where $\mathcal{I}_{\mathsf{TLS}}$ is the set of inliers selected by (TLS), $\boldsymbol{A}_{\mathcal{J}}$ is the matrix obtained by horizontally stacking all submatrices $\boldsymbol{A}_i$ for all $i \in \mathcal{J}$, and $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix. Moreover, if the inliers are noiseless, i.e., $\boldsymbol{\epsilon} = \boldsymbol{0}$ in eq. (17), and for a sufficiently small $\bar{c} > 0$, $\boldsymbol{x}_{\mathsf{TLS}} = \boldsymbol{x}^\circ$.*

Proof. Call $\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})$ the inlier set corresponding to an estimate $\boldsymbol{x}_{\mathsf{TLS}}$ (i.e., $\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}}) \triangleq \{i \in [n] : \left\| \boldsymbol{y}_i - \boldsymbol{A}^\mathsf{T} \boldsymbol{x}_{\mathsf{TLS}} \right\|_2 \le \bar{c}\}$). Moreover, define the TLS cost at $\boldsymbol{x}_{\mathsf{TLS}}$ as:

$$f(\boldsymbol{x}_{\mathsf{TLS}}) = \sum_{i \in \mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}_{\mathsf{TLS}} \right\|_2^2 + \bar{c}^2 (n - |\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})|). \tag{A73}$$

We first prove that $|\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})| \ge \alpha\, n - \frac{\gamma^\circ}{\bar{c}^2}$. Towards this goal, we observe that the cost evaluated at the ground truth $\boldsymbol{x}^\circ$ is:

$$f(\boldsymbol{x}^\circ) = \gamma^\circ + \bar{c}^2(n - |\mathcal{I}|) = \gamma^\circ + \bar{c}^2(n - \alpha\, n) \tag{A74}$$

which follows from the assumption that the inliers have squared residual error γ° and there are $\alpha\, n$ of them. Now assume by contradiction that there exists an $\boldsymbol{x}_{\mathsf{TLS}}$ that solves (TLS) and is such that $|\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})| < \alpha\, n - \frac{\gamma^\circ}{\bar{c}^2}$. Such an estimate would achieve a cost:

$$f(\boldsymbol{x}_{\mathsf{TLS}}) = \sum_{i \in \mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}_{\mathsf{TLS}} \right\|_2^2 + \bar{c}^2 (n - |\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})|) \tag{A75}$$

$$> \sum_{i \in \mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}_{\mathsf{TLS}} \right\|_2^2 + \bar{c}^2 \left(n - \alpha\, n + \frac{\gamma^\circ}{\bar{c}^2} \right) \tag{A76}$$

$$\ge \bar{c}^2 \left(n - \alpha\, n + \frac{\gamma^\circ}{\bar{c}^2} \right) = \gamma^\circ + \bar{c}^2 (n - \alpha\, n), \tag{A77}$$

which is larger than $f(\boldsymbol{x}^\circ)$, hence contradicting optimality of $\boldsymbol{x}_{\mathsf{TLS}}$, and implying $|\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})| \ge \alpha\, n - \frac{\gamma^\circ}{\bar{c}^2}$.

Since $|\mathcal{I}(\boldsymbol{x}^\circ)| = \alpha\, n$ and $|\mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})| \ge \alpha\, n - \frac{\gamma^\circ}{\bar{c}^2}$ then:

$$|\mathcal{I}(\boldsymbol{x}^\circ) \cap \mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})| \overset{\overbrace{\ge}^{\text{sets overlap in } [n]}}{} 2\alpha\, n - n - \frac{\gamma^\circ}{\bar{c}^2} \overset{\overbrace{\ge}^{\text{using } \alpha\, n \ge \frac{n+\bar{d}}{2} + \frac{\gamma^\circ}{\bar{c}^2}}}{} \bar{d}, \tag{A78}$$

The subset of measurements $|\mathcal{I}(\boldsymbol{x}^\circ) \cap \mathcal{I}(\boldsymbol{x}_{\mathsf{TLS}})|$ are simultaneously solved by $\boldsymbol{x}_{\mathsf{TLS}}$ and $\boldsymbol{x}^\circ$ (*i.e.*, are such that $\left\|\boldsymbol{y}_i - \boldsymbol{A}^\mathsf{T}\boldsymbol{x}\right\|_2 \leq \bar{c}$ for both $\boldsymbol{x} = \boldsymbol{x}^\circ$ and $\boldsymbol{x} = \boldsymbol{x}_{\mathsf{TLS}}$). Therefore, we can follow the same line of thoughts as in the proof of Proposition 5, and prove the first claim.

In the case of noiseless inliers, $\gamma^\circ = 0$ (or, equivalently, $\boldsymbol{y}_i - \boldsymbol{A}^\mathsf{T}\boldsymbol{x}^\circ = \boldsymbol{0}$, for all $i \in \mathcal{I}$) and we can always choose $\bar{c} > 0$ small enough such that the corresponding estimate $\boldsymbol{x}_{\mathsf{TLS}}$ satisfies the selected measurements exactly, *i.e.*, $\boldsymbol{y}_i - \boldsymbol{A}^\mathsf{T}\boldsymbol{x}_{\mathsf{TLS}} = \boldsymbol{0}$. Therefore, we can follow the same line of the proof of Proposition 5 (for the case of noiseless inliers) to conclude $\boldsymbol{x}_{\mathsf{TLS}} = \boldsymbol{x}^\circ$. ∎

E Proof of Theorem 12: Contract for Relaxation of (LTS1)

We start by restating the theorem for the reader's convenience.

Theorem A29 (Restatement of Theorem 12). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and known outlier rate β. Call $\ddot{\mathcal{I}}$ the set of uncorrupted measurements $(\boldsymbol{y}_i^\star, \boldsymbol{A}_i^\star)$, $i \in [n]$, where the outliers are replaced by inliers and assume that the set of matrices $\boldsymbol{A}_i^\star$, $i \in \ddot{\mathcal{I}}$, is k-certifiably C-hypercontractive with $k \geq 4$. Moreover, assume $\beta < \beta_{\max} = \sqrt[\frac{k}{2}-1]{1/(C(k/2)^{\frac{k}{2}}2^{3k-1})}$. Then, Algorithm 1 with relaxation order $r \geq k$ outputs an estimate $\boldsymbol{x}_{\mathsf{lts-sdp1}}$ (not necessarily in $\mathbb{X}$) such that:*

$$\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}_{\mathsf{lts-sdp1}}) \leq (1 + C_1(k,\beta)^{\frac{2}{k}}) \, \mathsf{opt}_{\ddot{\mathcal{I}}} + C_2(k,\beta)^{\frac{2}{k}} \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star \right\|_2^k \right)^{\frac{2}{k}}, \tag{A79}$$

where $C_1(k,\beta)$ and $C_2(k,\beta)$ are given functions, $\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x} \right\|_2^2$ is the residual error of an estimate $\boldsymbol{x}$ with respect to the inliers $\ddot{\mathcal{I}}$, $\boldsymbol{x}^\star \triangleq \arg\min_{\boldsymbol{x} \in \mathbb{X}} \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x} \right\|_2^2$ is the best estimate from an oracle estimator that has access to all the inliers, and $\mathsf{opt}_{\ddot{\mathcal{I}}} \triangleq \mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}^\star)$ is the corresponding residual error with respect to the inliers $\ddot{\mathcal{I}}$.

The proof is an adaptation of Lemma 5.3 and Lemma 5.6 in [15] to the case of vector-valued measurements. Let us start by clarifying all relevant notation:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}} \triangleq \left\{ \begin{array}{c} \omega_i^2 = \omega_i, \ \ i \in [n] \\ \sum_{i=1}^{n} \omega_i = \alpha n \\ \omega_i \cdot (\bar{\boldsymbol{y}}_i - \boldsymbol{y}_i) = \boldsymbol{0} \ \ i \in [n] \\ \omega_i \cdot (\bar{\boldsymbol{A}}_i - \boldsymbol{A}_i) = \boldsymbol{0} \ \ i \in [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\} \qquad \text{(constraints in (LTS1))} \tag{A80}$$

$$\{\boldsymbol{y}_i, \boldsymbol{A}_i\}_{i \in [n]} \qquad \text{(given measurements)} \tag{A81}$$

$$\{\boldsymbol{y}_i^\star \boldsymbol{A}_i^\star\}_{i \in [n]}, \qquad \text{(uncorrupted measurements with outliers replaced by inliers)} \tag{A82}$$

$$\boldsymbol{V} \triangleq \{\bar{\boldsymbol{y}}_i \bar{\boldsymbol{A}}_i\}_{i \in [n]}, \qquad \text{(auxiliary variables in (LTS1))} \tag{A83}$$

$$\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2 \quad \text{(error of } \boldsymbol{x} \text{ w.r.t. uncorrupted measurements)} \tag{A84}$$

$$\mathsf{err}(\boldsymbol{x},\boldsymbol{V}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left\|\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{A}}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2 \quad \text{(cost in (LTS1) without exponent } {}^k\!/{}_2). \tag{A85}$$

$$\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\omega_i'\left\|\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{A}}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2 \quad \text{(error w.r.t. set of measurements specified by } \boldsymbol{\omega}') \tag{A86}$$

Proof. The proof is quite involved and proceeds in two steps. First, we derive an sos proof that states that the inliers picked up by any feasible solution for (LTS1) must also satisfy a desired error bound. Then, we move to pseudo-expectations and conclude that the result of the moment relaxation must satisfy the same bound, which can be manipulated into eq. (A79).

Sos proof of robust certifiability (adapted from Lemma 5.6 in [15]). For a given tuple $(\boldsymbol{\omega},\boldsymbol{x},\boldsymbol{V})$ that satisfies $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$, let $\boldsymbol{\omega}'$ be such that $\omega_i' = \omega_i$ iff i is an inlier and $\omega_i' = 0$ otherwise (intuitively, $\boldsymbol{\omega}'$ is the indicator for the subset of the selected measurements $\boldsymbol{\omega}$ that are inliers).[38] Then call the corresponding error $\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})$ as in (A86). We note that:

$$\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}) = \frac{1}{n}\sum_{i=1}^{n}\omega_i'\left\|\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{A}}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\omega_i'\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2. \tag{A87}$$

The previous equality follows from the fact that (i) by definition, $\omega_i' = 1$ implies $\omega_i = 1$ and since the tuple $(\boldsymbol{\omega},\boldsymbol{x},\boldsymbol{V})$ satisfies $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$, then whenever $\omega_i = 1$ we must have $\bar{\boldsymbol{A}}_i = \boldsymbol{A}_i$ and $\bar{\boldsymbol{y}}_i = \boldsymbol{y}_i$, and (ii) ω_i' can only be 1 for inliers, for which $\boldsymbol{A}_i = \boldsymbol{A}_i^\star$ and $\boldsymbol{y}_i = \boldsymbol{y}_i^\star$. Therefore, $\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})$ is essentially the error attained by $\boldsymbol{x}$, but restricted to the true inliers in the set of measurements selected by $\boldsymbol{\omega}$.

We now show that any set of variables $(\boldsymbol{\omega},\boldsymbol{x},\boldsymbol{V})$ that are feasible for (LTS1) (*i.e.*, that satisfy the constraint set $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$), must also satisfy the following bound

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}} \left|\frac{k}{\boldsymbol{x}}\right. \left(\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}) - \mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right)^{\frac{k}{2}} \leq C_1(k,\beta)(\mathsf{err}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}} + C_2(k,\beta)\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k. \tag{A88}$$

We start by noting that $\sum_{i=1}^{n}\omega_i' \geq (1-2\beta)n$: this follows from the fact that the two sets, the selected measurements $\{i : \omega_i = 1\}$ and the set of true inliers, have each size $(1-\beta)n$, hence their intersection must contain at least $(1-2\beta)n$ measurements. Therefore:

$$\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i')^2 \overset{\text{binary variable}}{=\joinrel=} \frac{1}{n}\sum_{i=1}^{n}(1-\omega_i') = 1 - \frac{1}{n}\sum_{i=1}^{n}\omega_i' \leq 1 - (1-2\beta) \text{ hence:}$$

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}} \overset{\textit{Fact A13}}{\left|\frac{2}{\boldsymbol{\omega}'}\right.} \left\{\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i')^2 \leq 2\beta\right\}. \tag{A89}$$

Now we note that from definition (A84), we can expand $\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})$ as:

$$\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\omega_i'\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}(1-\omega_i')\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2 \tag{A90}$$

_____

[38]Note that $\boldsymbol{\omega}'$ is not required to (and typically does not) satisfy $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$.

Combining (A90) and (A87), and observing that the result is a sum of squares, we get:

$$\underbrace{\left|\tfrac{4}{\boldsymbol{\omega},\boldsymbol{x}}\right.\; \mathrm{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}) - \mathrm{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}) = \overbrace{\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i')\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}\right\|_2^2}^{\text{error for uncorrupted measurements not selected by }\boldsymbol{\omega}'}}_{} \qquad \geq 0 \qquad \text{(A91)}$$

Elevating both members of (A91) to $k/2$, and then using the sos version of Hölder's inequality (Fact A23), we get:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\left|\tfrac{k}{\boldsymbol{\omega},\boldsymbol{x},\boldsymbol{V}}\right.\left(\mathrm{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x}) - \mathrm{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right)^{\frac{k}{2}} \overset{\text{using (A91)}}{=} \left(\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i')\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}\right\|_2^2\right)^{\frac{k}{2}} \qquad \text{(A92)}$$

$$\leq \overbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i')^2\right)^{\frac{k}{2}-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}\right\|_2^k\right)}^{\text{using (A67) in Fact A23}} \overset{\text{using (A89)}}{\leq} (2\beta)^{\frac{k}{2}-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}\right\|_2^k\right). \qquad \text{(A93)}$$

Note that Fact A23 requires the exponent to be ≥ 1 and a power of 2, which in turns implies $\frac{k}{2}\geq 2$ or $k\geq 4$, as required by the statement of the theorem. Now we observe that:

$$\left|\tfrac{k}{\boldsymbol{x}}\right.\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}\right\|_2^k\right) \qquad \text{(A94)}$$

$$\overset{\text{adding/subtracting }(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}}{=} \left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}+(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^k\right) = \qquad \text{(A95)}$$

$$\overset{\text{using Fact A20}}{\leq} 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^k + 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{A}_i^{\star})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{x}^{\star})\right\|_2^k. \qquad \text{(A96)}$$

By certifiable hypercontractivity of $\boldsymbol{A}_i^{\star}$, $i\in[n]$:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\left|\tfrac{k}{\boldsymbol{x}}\right.\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{A}_i^{\star})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{x}^{\star})\right\|_2^k \leq C(k/2)^{\frac{k}{2}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{A}_i^{\star})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{x}^{\star})\right\|_2^2\right)^{\frac{k}{2}}. \qquad \text{(A97)}$$

We can further bound the term above as follows:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\left|\tfrac{k}{\boldsymbol{x}}\right.\left(\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{A}_i^{\star})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{x}^{\star})\right\|_2^2\right)^{\frac{k}{2}} \qquad \text{(A98)}$$

$$\overset{\text{adding/subtracting }\boldsymbol{y}_i^{\star}}{=} \left(\frac{1}{n}\sum_{i=1}^{n}\left\|-\boldsymbol{y}_i^{\star}+(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}+\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^2\right)^{\frac{k}{2}} \qquad \text{(A99)}$$

$$\overset{\text{using Fact A19}}{\leq} \left(\frac{1}{n}\sum_{i=1}^{n}\left(2\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}\right\|_2^2+2\left\|\boldsymbol{y}_i^{\star}-(\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^2\right)\right)^{\frac{k}{2}} \qquad \text{(A100)}$$

$$\overset{\text{rearranging}}{=} \left(2\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2 + 2\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^2\right)^{\frac{k}{2}} \tag{A101}$$

$$\overset{\text{using Fact A15}}{\leq} 2^{\frac{k}{2}}\left(2\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2\right)^{\frac{k}{2}} + 2^{\frac{k}{2}}\left(2\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^2\right)^{\frac{k}{2}} \tag{A102}$$

$$\overset{\text{rearranging}}{=} 2^k\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2\right)^{\frac{k}{2}} + 2^k\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^2\right)^{\frac{k}{2}}. \tag{A103}$$

Finally, using again the sos version of Hölder's inequality

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\left|\frac{k}{\boldsymbol{x}}\right. \left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^2\right)^{\frac{k}{2}} \leq \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k. \tag{A104}$$

Combining the above:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\left|\frac{k}{\boldsymbol{x}}\right. \left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^k\right) \tag{A105}$$

$$\overset{\text{(A96)}}{\leq} 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k + 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{A}_i^\star)^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}^\star)\right\|_2^k \tag{A106}$$

$$\overset{\text{(A97)}}{\leq} 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k + 2^k C(k/2)^{\frac{k}{2}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{A}_i^\star)^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}^\star)\right\|_2^2\right)^{\frac{k}{2}} \tag{A107}$$

$$\overset{\text{(A103)}}{\leq} 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k \tag{A108}$$

$$+2^k C(k/2)^{\frac{k}{2}}\left(2^k\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2\right)^{\frac{k}{2}} + 2^k\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^2\right)^{\frac{k}{2}}\right) \tag{A109}$$

$$\overset{\text{(A104)}}{\leq} 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k \tag{A110}$$

$$+2^k C(k/2)^{\frac{k}{2}}\left(2^k\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2\right)^{\frac{k}{2}} + 2^k\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k\right) \tag{A111}$$

$$\overset{\text{rearranging}}{=} C(k/2)^{\frac{k}{2}}\, 2^{2k}\overbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}\right\|_2^2\right)}^{\text{err}_{\mathcal{T}^i}(\boldsymbol{x})}{}^{\frac{k}{2}} \tag{A112}$$

$$+\left(2^k + C(k/2)^{\frac{k}{2}}\, 2^{2k}\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k. \tag{A113}$$

79

Hence, together with (A93):

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\Big|_{\boldsymbol{x}}^{k}\left(\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})-\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right)^{\frac{k}{2}}\le(2\beta)^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{2k}\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})^{\frac{k}{2}} \tag{A114}$$

$$+(2\beta)^{\frac{k}{2}-1}\left(2^{k}+C(k/2)^{\frac{k}{2}}\,2^{2k}\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star-(\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k. \tag{A115}$$

Applying Fact A17 to the right-hand-side with $a=\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})$, $b=\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})$, $\delta^{\frac{k}{2}}=(2\beta)^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{2k}$, and exponent $k/2$:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\Big|_{\boldsymbol{x}}^{k}\left(\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})-\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right)^{\frac{k}{2}}\le 2^{\frac{k}{2}}\overbrace{(2\beta)^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{2k}}^{=\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}2^{3k-1}}(\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})-\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}} \tag{A116}$$

$$+2^{\frac{k}{2}}(2\beta)^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{2k}(\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}} \tag{A117}$$

$$+(2\beta)^{\frac{k}{2}-1}\left(2^{k}+C(k/2)^{\frac{k}{2}}\,2^{2k}\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star-(\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k. \tag{A118}$$

Rearranging the terms:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\Big|_{\boldsymbol{x}}^{k}(1-\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{3k-1})(\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})-\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}}\le \beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{3k-1}(\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}}$$

$$+(2\beta)^{\frac{k}{2}-1}\left(2^{k}+C(k/2)^{\frac{k}{2}}\,2^{2k}\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star-(\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k. \tag{A119}$$

Noting that choosing $\beta<\sqrt[\frac{k}{2}-1]{\dfrac{1}{C(k/2)^{\frac{k}{2}}2^{3k-1}}}$ makes the constant $1-\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{3k-1}$ positive, we can divive both members of the inequality (A119) by such constant and obtain:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\Big|_{\boldsymbol{x}}^{k}\left(\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})-\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right)^{\frac{k}{2}}\le\frac{\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{3k-1}}{1-\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{3k-1}}(\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}} \tag{A120}$$

$$+\frac{(2\beta)^{\frac{k}{2}-1}\left(2^{k}+C(k/2)^{\frac{k}{2}}\,2^{2k}\right)}{1-\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}\,2^{3k-1}}\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star-(\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k, \tag{A121}$$

Now defining $C_1(k,\beta)\triangleq\dfrac{\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}2^{3k-1}}{1-\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}2^{3k-1}}$ and $C_2(k,\beta)\triangleq\dfrac{(2\beta)^{\frac{k}{2}-1}\left(2^{k}+C(k/2)^{\frac{k}{2}}2^{2k}\right)}{1-\beta^{\frac{k}{2}-1}C(k/2)^{\frac{k}{2}}2^{3k-1}}$, and noting that $\Big|_{\boldsymbol{x}}^{k}\mathsf{err}(\boldsymbol{x},\boldsymbol{V})\ge\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})$,[39] we finally get:

$$\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}\Big|_{\boldsymbol{x}}^{k}\left(\mathsf{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})-\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right)^{\frac{k}{2}}\le C_1(k,\beta)(\mathsf{err}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}}+C_2(k,\beta)\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^\star-(\boldsymbol{A}_i^\star)^\mathsf{T}\boldsymbol{x}^\star\right\|_2^k, \tag{A122}$$

which matches our claim in (A88).

---

[39]This follows from the definition of the two errors: $\mathsf{err}(\boldsymbol{x},\boldsymbol{V})-\mathsf{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})=\frac{1}{n}\sum_{i=1}^{n}\left\|\bar{\boldsymbol{y}}_i-\bar{\boldsymbol{A}}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2-\frac{1}{n}\sum_{i=1}^{n}\omega_i'\left\|\bar{\boldsymbol{y}}_i-\bar{\boldsymbol{A}}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2=\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i')\left\|\bar{\boldsymbol{y}}_i-\bar{\boldsymbol{A}}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2$, which is a sum of squares.

Completing the proof by moving to pseudo-distributions. Consider a pseudo-distribution $\tilde{\mu}$ that satisfies $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$. Using the sos proof in (A88) and thanks to Fact A10, we conclude that if $\tilde{\mu}$ satisfies $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$ then it must also satisfy:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[(\text{err}_{\ddot{\mathcal{I}}(\boldsymbol{x})} - \text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}}\right] \leq C_1(k,\beta) \overbrace{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}(\boldsymbol{x},\boldsymbol{V})^{\frac{k}{2}}\right]}^{\text{by definition this is } \widehat{\text{opt}}_{\text{lts}-\text{sdp1}}^{\frac{k}{2}}} + C_2(k,\beta)\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star} - (\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^{k}\right).$$

(A123)

Elevating to the power $\frac{2}{k}$ both sides and recalling that $(a+b)^q \leq a^q + b^q$ for any $0 < q < 1$:

$$\left(\tilde{\mathbb{E}}_{\tilde{\mu}}\left[(\text{err}_{\ddot{\mathcal{I}}(\boldsymbol{x})} - \text{err}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}}\right]\right)^{\frac{2}{k}} \leq C_1(k,\beta)^{\frac{2}{k}}\,\widehat{\text{opt}}_{\text{lts}-\text{sdp1}} + C_2(k,\beta)^{\frac{2}{k}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star} - (\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^{k}\right)^{\frac{2}{k}}.$$

(A124)

Now using the sos version of Hölder's inequality for pseudo-expectations (Fact A7, eq. (A17)):

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}_{\ddot{\mathcal{I}}(\boldsymbol{x})} - \text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right]^{\frac{k}{2}} \leq \tilde{\mathbb{E}}_{\tilde{\mu}}\left[(\text{err}_{\ddot{\mathcal{I}}(\boldsymbol{x})} - \text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))^{\frac{k}{2}}\right],$$

(A125)

and therefore (A124) becomes:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[(\text{err}_{\ddot{\mathcal{I}}(\boldsymbol{x})} - \text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}))\right] \leq C_1(k,\beta)^{\frac{2}{k}}\,\widehat{\text{opt}}_{\text{lts}-\text{sdp1}} + C_2(k,\beta)^{\frac{2}{k}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star} - (\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^{k}\right)^{\frac{2}{k}}.$$

(A126)

By linearity of the pseudo-expectation and rearranging:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}_{\ddot{\mathcal{I}}(\boldsymbol{x})}\right] \leq \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right] + C_1(k,\beta)^{\frac{2}{k}}\,\widehat{\text{opt}}_{\text{lts}-\text{sdp1}} + C_2(k,\beta)^{\frac{2}{k}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star} - (\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^{k}\right)^{\frac{2}{k}}.$$

(A127)

Noting that $\left|\frac{4}{\boldsymbol{\omega},\boldsymbol{x}}\right.\{\text{err}(\boldsymbol{x},\boldsymbol{V}) - \text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V}) \geq 0\}$,[40] and using Fact A10, we get $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right] \leq \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}(\boldsymbol{x},\boldsymbol{V})\right]$. Moreover, we observe

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}(\boldsymbol{x},\boldsymbol{V})\right] = \left(\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}(\boldsymbol{x},\boldsymbol{V})\right]^{\frac{k}{2}}\right)^{\frac{2}{k}} \overset{\text{Fact A7, eq. (A17)}}{\leq} \left(\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}(\boldsymbol{x},\boldsymbol{V})^{\frac{k}{2}}\right]\right)^{\frac{2}{k}} = \widehat{\text{opt}}_{\text{lts}-\text{sdp1}}$$

(A128)

concluding that $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})\right] \leq \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}(\boldsymbol{x},\boldsymbol{V})\right] \leq \widehat{\text{opt}}_{\text{lts}-\text{sdp1}}$. Using this inequality in (A127):

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\text{err}_{\ddot{\mathcal{I}}(\boldsymbol{x})}\right] \leq (1 + C_1(k,\beta)^{\frac{2}{k}})\,\widehat{\text{opt}}_{\text{lts}-\text{sdp1}} + C_2(k,\beta)^{\frac{2}{k}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{\star} - (\boldsymbol{A}_i^{\star})^{\mathsf{T}}\boldsymbol{x}^{\star}\right\|_2^{k}\right)^{\frac{2}{k}}$$

(A129)

---

[40]This again follows from the definition of the errors $\text{err}(\boldsymbol{x},\boldsymbol{V})$ and $\text{err}_{\boldsymbol{\omega}'}(\boldsymbol{x},\boldsymbol{V})$, whose difference is a sum of squares.

Applying Hölder's inequality (Fact A7, eq. (A17)) one last time, we get $\text{err}_{\ddot{\mathcal{I}}}(\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{x}]) \leq \tilde{\mathbb{E}}_{\tilde{\mu}}[\text{err}_{\ddot{\mathcal{I}}}(\boldsymbol{x})]$, which leads to:

$$\text{err}_{\ddot{\mathcal{I}}}(\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{x}]) \leq (1 + C_1(k,\beta)^{\frac{2}{k}})\,\widehat{\text{opt}}_{\text{lts-sdp1}} + C_2(k,\beta)^{\frac{2}{k}} \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T} \boldsymbol{x}^\star \right\|_2^k \right)^{\frac{2}{k}}. \qquad (A130)$$

Finally, we need to prove that $\widehat{\text{opt}}_{\text{lts-sdp1}} \leq \text{opt}_{\ddot{\mathcal{I}}}$. Towards this goal, we observe that the (pseudo-)distribution supported on the point $(\boldsymbol{\omega}^\star, \boldsymbol{x}^\star, \boldsymbol{V})$ where $\omega_i^\star = 1$ for the true inliers and zero otherwise is feasible for $\mathcal{L}_{\boldsymbol{\omega},\boldsymbol{x}}$, hence by optimality $\widehat{\text{opt}}_{\text{lts-sdp1}}^{\frac{k}{2}} \leq \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \bar{\boldsymbol{y}}_i - \bar{\boldsymbol{A}}_i^\mathsf{T} \boldsymbol{x}^\star \right\|_2^2 \right)^{\frac{k}{2}}$, from which it follows:

$$\widehat{\text{opt}}_{\text{lts-sdp1}} \leq \frac{1}{n} \sum_{i=1}^{n} \left\| \bar{\boldsymbol{y}}_i - \bar{\boldsymbol{A}}_i^\mathsf{T} \boldsymbol{x}^\star \right\|_2^2 \overset{\overset{\bar{\boldsymbol{y}}_i, \bar{\boldsymbol{A}}_i \text{ are inliers or zero}}{\frown}}{\leq} \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T} \boldsymbol{x}^\star \right\|_2 = \text{opt}_{\ddot{\mathcal{I}}}. \qquad (A131)$$

Substituting (A131) back into (A130):

$$\text{err}_{\ddot{\mathcal{I}}}(\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{x}]) \leq (1 + C_1(k,\beta)^{\frac{2}{k}})\,\text{opt}_{\ddot{\mathcal{I}}} + C_2(k,\beta)^{\frac{2}{k}} \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i^\star - (\boldsymbol{A}_i^\star)^\mathsf{T} \boldsymbol{x}^\star \right\|_2^k \right)^{\frac{2}{k}}, \qquad (A132)$$

which proves the claim of Theorem 12.

∎

F Proof of Proposition 13: Contract for Relaxation of (LTS2)

We start by restating the proposition for the reader's convenience.

Proposition A30 (Restatement of Proposition 13). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and outlier rate $\beta < 0.5$ (or, equivalently, inlier rate $\alpha = 1 - \beta > 0.5$). Call $\mathcal{I}$ the set of inliers and assume that the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2 \eta^2 (1-2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated for some $\eta > 0$. Then, Algorithm 2 with relaxation order $r \geq k/2$ outputs an estimate $\boldsymbol{x}_{\text{lts-sdp2}}$ (not necessarily in $\mathbb{X}$) such that:*

$$\left\| \boldsymbol{x}_{\text{lts-sdp2}} - \boldsymbol{x}^\circ \right\|_2 \leq M_x \left(\frac{\alpha\,\eta}{2} + 2\frac{1-\alpha}{\alpha} \right). \qquad (A133)$$

Towards proving the proposition we need to prove two technical lemmas (Lemmas A31 and A32 below). These lemmas extend results [37] to vector-valued and noisy measurements, and while in [37] they have been proposed to attack the high-outlier case (*i.e.*, for list-decodable regression), we show they are also useful to prove estimation contracts for the low-outlier case.

Note that the two lemmas below use a subset of constraints compared to the one in the constraint set of (LTS2) (*i.e.*, the set $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$ below does not contain the constraint $\sum_{i=1}^{n} \omega_i = \alpha n$): this will allow us to use them also to discuss the performance of (MC) and (TLS) later on.

Lemma A31 (Adapted from Lemma 4.1 in [37]). *Consider the following constraint set, for given measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, a constant $\bar{c} \geq 0$, and where $\mathbb{X}$ is an explicitly bounded basic*

*semi-algebraic set (*cf. *Assumption 2):*

$$\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \doteq \left\{ \begin{array}{c} \omega_i^2 = \omega_i, \ \ i \in [n] \\ \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x} \right\|_2^2 \leq \bar{c}^2 \ \ i \in [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\}. \tag{A134}$$

For any $t \geq k$ and set of n measurements with at least $\alpha\,n$ inliers, such that for the set of inliers $\mathcal{I}$, the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2\eta^2(1-2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated,[41]*

$$\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \left|\frac{t}{\boldsymbol{\omega},\boldsymbol{x}} \right. \left\{ \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \omega_i \left\| \boldsymbol{x} - \boldsymbol{x}^\circ \right\|_2^2 \leq \frac{\alpha^2 \eta^2 M_x^2}{4} \right\}. \tag{A135}$$

Proof. We follow the same logic as the proof of Lemma 4.1 in [37], but provide a slightly simpler derivation, based on our definition of certifiable anti-concentration. We first observe that for the inliers (*i.e.,* $i \in \mathcal{I}$) it holds:[42]

$$\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \left|\frac{t}{\boldsymbol{x}} \right. \omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2^2 \overset{\overbrace{\text{adding/subtracting } \boldsymbol{y}_i}}{=} \omega_i \cdot \left\| (\boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}^\circ) - (\boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}) \right\|_2^2 \tag{A136}$$

$$\overset{\overbrace{\text{Fact A19}}}{\leq} \omega_i \cdot \left(2 \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}^\circ \right\|_2^2 + 2 \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x} \right\|_2^2 \right) \tag{A137}$$

$$\overset{\overbrace{\omega_i \leq 1 \text{ and Fact A13}}}{\leq} 2 \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}^\circ \right\|_2^2 + 2 \cdot \omega_i \cdot \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x} \right\|_2^2 \tag{A138}$$

(since $(\boldsymbol{\omega}, \boldsymbol{x})$ satisfy $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$, and inliers by definition satisfy $\left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}^\circ \right\|_2^2 \leq \bar{c}^2$)

$$\leq 4\bar{c}^2. \tag{A139}$$

Since the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2\eta^2(1-2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated, then there exists a univariate polynomial p such that for every $i \in \mathcal{I}$ and for every $t \geq k$:

$$\overbrace{\left\{ \omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2^2 \leq 4\bar{c}^2 \right\}}^{\text{from (A139)}} \overbrace{\left|\frac{t}{\boldsymbol{x}} \right. p^2 \left(\omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \geq (1 - 2\bar{c})^2}^{\text{using (32) with } \delta = 2\bar{c}} \tag{A140}$$

and

$$\left\| \boldsymbol{x} \right\|_2^2 \leq M_x \overbrace{\left|\frac{t}{\boldsymbol{x}} \right. \left\| \boldsymbol{x} - \boldsymbol{x}^\circ \right\|_2^2 \leq 4M_x^2}^{\text{Fact A19}} \tag{A141}$$

$$\overbrace{\left|\frac{t}{\boldsymbol{x}} \right. \left\{ \left\| \boldsymbol{x} - \boldsymbol{x}^\circ \right\|_2^2 \cdot \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p^2 \left(\left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \leq \frac{\alpha^2 \eta^2 (1 - 2\bar{c})^2 M_x^2}{4} \right\}}^{\text{using certifiable anti-concentration in eq. (33)}}. \tag{A142}$$

---

[41] The constant "32" in the anti-concentration requirement is arbitrary (*i.e.,* it just amounts to a re-scaling of the parameter η) and has been chosen to keep the result consistent with the original statement in [37].

[42] Observe the analogy with the proof of Proposition 5.

Now we note that:

$$\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \mid \frac{t}{\boldsymbol{\omega},\boldsymbol{x}} \, \omega_i \cdot p^2 \left(\omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \overset{\omega_i^2 = \omega_i}{=} \left(\omega_i \cdot p \left(\omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \right)^2 \quad \text{(A143)}$$

$$= \overbrace{\left(\omega_i \cdot \left(1 + h \left(\omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \right) \right)^2}^{\text{calling } h \text{ the homogeneous part of } p \text{ and since } p(0)=1} = \left(\omega_i + \omega_i \cdot h \left(\omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \right)^2 \quad \text{(A144)}$$

$$\overset{\omega_i^2 = \omega_i}{=} \left(\omega_i + \omega_i \cdot h \left(\left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \right)^2 = \omega_i^2 \cdot \left(1 + h \left(\left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \right)^2 \quad \text{(A145)}$$

$$= \omega_i^2 \cdot p^2 \left(\left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right) \overset{\omega_i \leq 1}{\leq} p^2 \left(\left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right). \quad \text{(A146)}$$

Combining the conclusions in (A140), (A142), and (A146) we obtain:

$$\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \mid \frac{t}{\boldsymbol{\omega},\boldsymbol{x}} \, \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \omega_i \left\| \boldsymbol{x} - \boldsymbol{x}^\circ \right\|_2^2 \leq \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \omega_i \left\| \boldsymbol{x} - \boldsymbol{x}^\circ \right\|_2^2 \cdot \overbrace{\frac{1}{(1-2\bar{c})^2} p^2 \left(\omega_i \cdot \left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right)}^{\text{since } \frac{1}{(1-2\bar{c})^2} p^2 \left(\omega_i \cdot \| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \|_2 \right) \geq 1 \text{ per (A140)}} \quad \text{(A147)}$$

$$\leq \overbrace{\frac{1}{(1-2\bar{c})^2} \left\| \boldsymbol{x} - \boldsymbol{x}^\circ \right\|_2^2 \cdot \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p^2 \left(\left\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \right\|_2 \right)}^{\text{using } \omega_i \cdot p^2 \left(\omega_i \cdot \| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \|_2 \right) \leq p^2 \left(\| \boldsymbol{A}_i^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^\circ) \|_2 \right) \text{ from (A146)}} \overset{\text{using (A142)}}{\leq} \frac{\alpha^2 \eta^2 M_x^2}{4}, \quad \text{(A148)}$$

which concludes the proof. ∎

Lemma A32 (Adapted from Lemma 4.2 in [37]). *Under the same assumptions of Lemma A31, for any pseudo-distribution $\tilde{\mu}$ of level at least k satisfying $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$,*

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 \leq \frac{\alpha \eta M_x}{2}, \quad \text{(A149)}$$

where the vectors $\boldsymbol{v}_i$ are extracted from the pseudo-moment matrix by setting $\boldsymbol{v}_i = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i \boldsymbol{x}]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}$ if $\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] > 0$, or $\boldsymbol{v}_i = \boldsymbol{0}$ otherwise, for $i \in [n]$.

Proof. By Lemma A31, we have $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \mid \frac{k}{\boldsymbol{\omega},\boldsymbol{x}} \left\{ \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \omega_i \left\| \boldsymbol{x} - \boldsymbol{x}^\circ \right\|_2^2 \leq \frac{\alpha^2 \eta^2 M_x^2}{4} \right\}$. We also have: $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \mid \frac{2}{\boldsymbol{\omega}} \left\{ \omega_i^2 = \omega_i \right\}$ for any i. Therefore:

$$\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}} \mid \frac{k}{\boldsymbol{\omega},\boldsymbol{x}} \left\{ \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left\| \omega_i \boldsymbol{x} - \omega_i \boldsymbol{x}^\circ \right\|_2^2 \leq \frac{\alpha^2 \eta^2 M_x^2}{4} \right\}. \quad \text{(A150)}$$

Since $\tilde{\mu}$ satisfies $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$, then it also satisfies:

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\left\| \omega_i \boldsymbol{x} - \omega_i \boldsymbol{x}^\circ \right\|_2^2 \right] \leq \frac{\alpha^2 \eta^2 M_x^2}{4}. \quad \text{(A151)}$$

Using the norm inequality for pseudo-distributions in Fact A8, we get $\left\| \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \boldsymbol{x} - \omega_i \boldsymbol{x}^\circ \right] \right\|_2^2 \leq$ $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[\left\| \omega_i \boldsymbol{x} - \omega_i \boldsymbol{x}^\circ \right\|_2^2 \right]$; then observing that for any m-vector $\boldsymbol{z}$, $\|\boldsymbol{z}\|_1 \leq \sqrt{m} \|\boldsymbol{z}\|_2$ or, equivalently, $\|\boldsymbol{z}\|_1^2 \leq m \|\boldsymbol{z}\|_2^2$ (below we will apply this inequality to the vector of size $|\mathcal{I}|$ with entries $z_i = \frac{1}{|\mathcal{I}|} \left\| \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \boldsymbol{x} \right] - \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \boldsymbol{x}^\circ \right\|_2$), and chaining the inequalities back to (A151):

$$
\overbrace{\left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left\| \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \boldsymbol{x} \right] - \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \boldsymbol{x}^\circ \right\|_2 \right)^2}^{\|\boldsymbol{z}\|_1^2} \overset{\|\boldsymbol{z}\|_1^2 \leq m\|\boldsymbol{z}\|_2^2}{\leq} \overbrace{\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left\| \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \boldsymbol{x} \right] - \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \boldsymbol{x}^\circ \right\|_2^2}^{m\|\boldsymbol{z}\|_2^2 \text{ with } m=|\mathcal{I}|} \tag{A152}
$$

$$
\overset{Fact\ A8}{\leq} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\left\| \omega_i \boldsymbol{x} - \omega_i \boldsymbol{x}^\circ \right\|_2^2 \right] \overset{(A151)}{\leq} \frac{\alpha^2 \eta^2 M_x^2}{4}. \tag{A153}
$$

Remembering that $\boldsymbol{v}_i = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i \boldsymbol{x}]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}$ if $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] > 0$, or $\boldsymbol{v}_i = \boldsymbol{0}$ otherwise, and taking the square root of both members in (A153):

$$
\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}, \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]>0} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 \overset{by\ def.\ of\ \boldsymbol{v}_i}{=} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left\| \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \boldsymbol{x} \right] - \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \boldsymbol{x}^\circ \right\|_2 \overset{(A153)}{\leq} \frac{\alpha \eta M_x}{2}, \tag{A154}
$$

concluding the proof of Lemma A32. ∎

Proof of Proposition 13: First of all, we note that since $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ in Algorithm 2 contains a superset of the constraints in $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$ defined in Lemma A31, the conclusions of Lemma A32 and Lemma A31 still hold if we replace $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$ with $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$. Therefore we have that any pseudo-distribution of level at least k (hence produced by a relaxation of order at least $k/2$) satisfying $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ also satisfies:

$$
\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 \leq \frac{\alpha \eta M_x}{2} \overset{|\mathcal{I}|=\alpha\, n}{\Longleftrightarrow} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 \leq \frac{\alpha^2 \eta\, M_x\, n}{2}. \tag{A155}
$$

Let us define the set of outliers $\mathcal{O} \triangleq [n] \backslash \mathcal{I}$. We observe that since $\tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \leq 1$, then $\sum_{i \in \mathcal{O}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \leq (1-\alpha)n$. Moreover, using the triangle inequality $\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq 2M_x$, hence:

$$
\sum_{i \in \mathcal{O}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 \leq 2nM_x(1-\alpha). \tag{A156}
$$

Using Eq. (A155) and Eq. (A156):

$$
\sum_{i=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 = \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 + \sum_{i \in \mathcal{O}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^\circ \right\|_2 \tag{A157}
$$

$$
\leq \frac{\alpha^2\, \eta\, M_x\, n}{2} + 2nM_x(1-\alpha). \tag{A158}
$$

Now note that any pseudo-distribution $\tilde{\mu}$ satisfying $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ is such that $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{i=1}^n \omega_i\right] = \alpha\,n$ (due to the constraint $\sum_{i=1}^n \omega_i = \alpha\,n$ in $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$), hence by linearity $\sum_{i=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right] = \alpha\,n$. Dividing both members of (A158) by $\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_j\right]$ (where we switched to using j as an index to avoid confusion):

$$\sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_j\right]} \left\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq \frac{1}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_j\right]} \left(\frac{\alpha^2\,\eta\,M_x\,n}{2} + 2nM_x(1-\alpha)\right) \tag{A159}$$

$$= \frac{1}{\alpha\,n} \left(\frac{\alpha^2\,\eta\,M_x\,n}{2} + 2nM_x(1-\alpha)\right). \tag{A160}$$

Using Jensen's inequality, we observe $\left\|\sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]}\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq \sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]}\left\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2$, hence (A180) becomes:

$$\left\|\sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_j\right]}\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq \frac{\alpha\,\eta\,M_x}{2} + 2M_x\frac{1-\alpha}{\alpha}, \tag{A161}$$

which, recalling that $\boldsymbol{x}_{\mathsf{lts-sdp2}} = \sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]}\boldsymbol{v}_i$, concludes the proof.

∎

G   Proof of Proposition 14: Contract for Relaxation of (MC1)

We start by restating the proposition for the reader's convenience.

Proposition A33 (Restatement of Proposition 14). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and outlier rate $\beta < 0.5$ (or, equivalently, inlier rate $\alpha = 1 - \beta > 0.5$). Call $\mathcal{I}$ the set of inliers and assume that the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2\eta^2(1-2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated for some $\eta > 0$. Then, Algorithm 3 with relaxation order $r \geq k/2$ outputs an estimate $\boldsymbol{x}_{\mathsf{mc-sdp}}$ (not necessarily in $\mathbb{X}$) such that:*

$$\left\|\boldsymbol{x}_{\mathsf{mc-sdp}} - \boldsymbol{x}^\circ\right\|_2 \leq M_x \left(\frac{\alpha\,\eta}{2} + 2\frac{1-\alpha}{\alpha}\right). \tag{A162}$$

Proof. First of all, we note that the constraint set $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$ in (MC1) is the same as Lemma A31 and Lemma A32. Therefore we have that any pseudo-distribution $\tilde{\mu}$ of level at least k (hence produced by a relaxation of order at least $k/2$) satisfying $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$ also satisfies:

$$\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]\left\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq \frac{\alpha\eta M_x}{2} \overset{|\mathcal{I}|=\alpha\,n}{\Longleftrightarrow} \sum_{i \in \mathcal{I}}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]\left\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq \frac{\alpha^2\,\eta\,M_x\,n}{2}. \tag{A163}$$

Let us define the set of outliers $\mathcal{O} \triangleq [n]\backslash\mathcal{I}$. We observe that since $\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right] \leq 1$, then $\sum_{i \in \mathcal{O}}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right] \leq (1-\alpha)n$. Moreover, using the triangle inequality $\left\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq 2M_x$, hence:

$$\sum_{i \in \mathcal{O}}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]\left\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq 2nM_x(1-\alpha). \tag{A164}$$

Using Eq. (A163) and Eq. (A164):

$$\sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 = \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 + \sum_{i \in \mathcal{O}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 \tag{A165}$$

$$\leq \frac{\alpha^2 \, \eta \, M_x \, n}{2} + 2n M_x (1 - \alpha). \tag{A166}$$

Let us call $\tilde{\mu}$ the pseudo-distribution that achieves the optimal solution in (MC1), and observe that the corresponding optimal objective $\sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \geq \alpha \, n$: this follows from optimality of $\tilde{\mu}$ and from the fact that the pseudo-distribution supported on the single point $(\boldsymbol{x}^{\circ}, \boldsymbol{\omega}^{\circ})$, where $\omega_i^{\circ} = 1$ if $i \in \mathcal{I}$ or zero otherwise, is feasible for (MC1) and achieves an objective $\alpha \, n$.

Now dividing both members of (A174) by $\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_j \right]$:

$$\sum_{i=1}^{n} \frac{\tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right]}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_j \right]} \left\| \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 \leq \frac{1}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_j \right]} \left(\frac{\alpha^2 \, \eta \, M_x \, n}{2} + 2n M_x (1 - \alpha) \right) \tag{A167}$$

$$\leq \frac{1}{\alpha \, n} \left(\frac{\alpha^2 \, \eta \, M_x \, n}{2} + 2n M_x (1 - \alpha) \right). \tag{A168}$$

Using Jensen's inequality $\left\| \sum_{i=1}^{n} \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 \leq \sum_{i=1}^{n} \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \left\| \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2$ hence (A168) becomes:

$$\left\| \sum_{i=1}^{n} \frac{\tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right]}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_j \right]} \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 \leq \frac{\alpha \, \eta \, M_x}{2} + 2 M_x \frac{1 - \alpha}{\alpha}, \tag{A169}$$

which, recalling that $\boldsymbol{x}_{\mathsf{mc-sdp}} = \sum_{i=1}^{n} \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \boldsymbol{v}_i$, concludes the proof. $\blacksquare$

H   Proof of Proposition 15: Contract for Relaxation of (TLS1)

We start by restating the proposition for the reader's convenience.

Proposition A34 (Restatement of Proposition 15)**.** *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and outlier rate $\beta < 0.5$ (or, equivalently, inlier rate $\alpha = 1 - \beta > 0.5$). Call $\mathcal{I}$ the set of inliers and assume that the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2 \eta^2 (1 - 2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated for some $\eta > 0$. Then, Algorithm 4 with relaxation order $r \geq k/2$ outputs an estimate $\boldsymbol{x}_{\mathsf{tls-sdp}}$ (not necessarily in $\mathbb{X}$) such that:*

$$\left\| \boldsymbol{x}_{\mathsf{tls-sdp}} - \boldsymbol{x}^{\circ} \right\|_2 \leq \frac{\alpha M_x n}{\alpha \, n - \frac{\gamma^{\circ}}{\bar{c}^2}} \left(\frac{\alpha \, \eta}{2} + 2 \frac{1 - \alpha}{\alpha} \right), \tag{A170}$$

where $\gamma^{\circ} \triangleq \sum_{i \in \mathcal{I}} \left\| \boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}^{\circ} \right\|_2^2$ is the squared residual error of the ground truth $\boldsymbol{x}^{\circ}$ over the inliers $\mathcal{I}$.

Proof. First of all, we note that the constraint set $\mathcal{M}_{\boldsymbol{\omega}, \boldsymbol{x}}$ in (TLS1) is the same as Lemma A31 and Lemma A32. Therefore we have that any pseudo-distribution $\tilde{\mu}$ of level at least k (hence produced by a relaxation of order at least $k/2$) satisfying $\mathcal{M}_{\boldsymbol{\omega}, \boldsymbol{x}}$ also satisfies:

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 \leq \frac{\alpha \eta M_x}{2} \overset{|\mathcal{I}| = \alpha \, n}{\Longleftrightarrow} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}} \left[\omega_i \right] \left\| \boldsymbol{v}_i - \boldsymbol{x}^{\circ} \right\|_2 \leq \frac{\alpha^2 \, \eta \, M_x \, n}{2}. \tag{A171}$$

Let us define the set of outliers $\mathcal{O} \triangleq [n] \backslash \mathcal{I}$. We observe that since $\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \leq 1$, then $\sum_{i \in \mathcal{O}} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \leq (1-\alpha)n$. Moreover, using the triangle inequality $\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq 2M_x$, hence:

$$\sum_{i \in \mathcal{O}} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq 2nM_x(1-\alpha). \tag{A172}$$

Using Eq. (A171) and Eq. (A172):

$$\sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 = \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 + \sum_{i \in \mathcal{O}} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \tag{A173}$$

$$\leq \frac{\alpha^2 \, \eta \, M_x \, n}{2} + 2nM_x(1-\alpha). \tag{A174}$$

Let us call $\tilde{\mu}$ the pseudo-distribution that achieves the optimal solution in (TLS1), and observe that $\tilde{\mu}$ achieves a cost:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\sum_{i=1}^{n} \omega_i \cdot \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}\right\|_2^2 + (1-\omega_i) \cdot \bar{c}^2\right] = \sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i \cdot \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}\right\|_2^2\right] + \sum_{i=1}^{n}(1 - \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]) \cdot \bar{c}^2. \tag{A175}$$

Now observe that the pseudo-distribution supported on the single point $(\boldsymbol{x}^\circ, \boldsymbol{\omega}^\circ)$, where $\omega_i^\circ = 1$ if $i \in \mathcal{I}$ or zero otherwise, is feasible for (TLS1) and achieves an objective $\sum_{i \in \mathcal{I}} \|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ\|_2^2 + (1-\alpha)n\bar{c}^2$. Therefore, by using (A175) and by optimality of $\tilde{\mu}$:

$$\sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i \cdot \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}\right\|_2^2\right] + \overbrace{\sum_{i=1}^{n}(1 - \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]) \cdot \bar{c}^2}^{=n\bar{c}^2 - \sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]\bar{c}^2} \leq \sum_{i \in \mathcal{I}} \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ\right\|_2^2 + \overbrace{(1-\alpha)n\bar{c}^2}^{=n\bar{c}^2 - \alpha \, n \, \bar{c}^2}. \tag{A176}$$

Rearranging the terms in the previous inequality:

$$\sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \geq \frac{1}{\bar{c}^2}\left(\sum_{i=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i \cdot \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}\right\|_2^2\right] - \sum_{i \in \mathcal{I}} \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ\right\|_2^2 + \alpha \, n \, \bar{c}^2\right) \tag{A177}$$

$$\geq \frac{1}{\bar{c}^2}\left(\alpha \, n \, \bar{c}^2 - \sum_{i \in \mathcal{I}} \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ\right\|_2^2\right) = \alpha \, n - \frac{1}{\bar{c}^2} \sum_{i \in \mathcal{I}} \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ\right\|_2^2. \tag{A178}$$

Now dividing both members of (A174) by $\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]$ and defining $\gamma^\circ \triangleq \sum_{i \in \mathcal{I}} \|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T} \boldsymbol{x}^\circ\|_2^2$:

$$\sum_{i=1}^{n} \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \frac{1}{\sum_{j=1}^{n} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]}\left(\frac{\alpha^2 \, \eta \, M_x \, n}{2} + 2nM_x(1-\alpha)\right) \tag{A179}$$

$$\overset{\text{using (A178)}}{\leq} \frac{1}{\alpha \, n - \frac{\gamma^\circ}{\bar{c}^2}}\left(\frac{\alpha^2 \, \eta \, M_x \, n}{2} + 2nM_x(1-\alpha)\right). \tag{A180}$$

88

Using Jensen's inequality $\left\|\sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq \sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2$ hence (A180) becomes:

$$\left\|\sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \boldsymbol{v}_i - \boldsymbol{x}^\circ\right\|_2 \leq \frac{1}{\alpha\, n - \frac{\gamma^\circ}{\bar{c}^2}}\left(\frac{\alpha^2\, \eta\, M_x\, n}{2} + 2nM_x(1-\alpha)\right), \qquad \text{(A181)}$$

which, recalling that $\boldsymbol{x}_{\mathsf{tls-sdp}} = \sum_{i=1}^n \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i]}{\sum_{j=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_j]} \boldsymbol{v}_i$, concludes the proof. ∎

I Proof of Theorem 17: Contract for Relaxation of (LDR)

We start by restating the theorem for the reader's convenience.

Theorem A35 (Restatement of Proposition 17). *Consider Problem 1 with measurements $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, and known outlier rate β (or, equivalently, known inlier rate $\alpha = 1 - \beta$), possibly with $\beta > 0.5$. Call $\mathcal{I}$ the set of inliers and assume that the set of matrices $\boldsymbol{A}_i$, $i \in \mathcal{I}$, is k-certifiably $(\frac{\alpha^2\eta^2(1-2\bar{c})^2}{32\bar{c}}, 2\bar{c}, 2M_x)$-anti-concentrated for some $\eta > 0$. Then, with probability at least $1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{N}{\alpha}}$ (over the draw of the samples in the algorithms), where $N \geq 1$ is a user-defined parameter, Algorithm 5 with relaxation order $r \geq k/2$ outputs a list $\mathcal{L}$ of size N/α such that there is an estimate $\boldsymbol{x} \in \mathcal{L}$ (with $\boldsymbol{x}$ not necessarily in $\mathbb{X}$) such that*

$$\|\boldsymbol{x} - \boldsymbol{x}^\circ\|_2 \leq \eta M_x. \qquad \text{(A182)}$$

Moreover, when $\alpha \geq 0.01$ (i.e., at least 1% of the measurements are inliers) and $N = 10$, the relation $\|\boldsymbol{x} - \boldsymbol{x}^\circ\|_2 \leq \eta M_x$ holds with probability at least 0.99 over the draw of the samples.

Proof. Note that Lemma A31 and Lemma A32 use a subset of constraints compared to the set $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ in (LDR) (*i.e.*, the set $\mathcal{M}_{\boldsymbol{\omega},\boldsymbol{x}}$ in the lemmas does not contain the constraint $\sum_{i=1}^n \omega_i = \alpha n$, while $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ does). Therefore, their conclusions will still hold in the context of (LDR). We start by proving the following lemma, which shows that the pseudo-distribution $\tilde{\mu}$ built by optimizing the moment relaxation of (LDR) "spreads" (*i.e.*, has enough support) across the inliers. The proof is an extension of Lemma 4.3 in [37] to the case of vector-valued measurements.

Lemma A36 (Adapted from Lemma 4.3 in [37]). *For any pseudo-distribution $\tilde{\mu}$ satisfying $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ that minimizes $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{\omega}]\right\|_2^2$, $\sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}}[\omega_i] \geq \alpha^2 n$.*

Proof. Let $\boldsymbol{u} = \frac{1}{\alpha n}\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{\omega}]$. Then, $\boldsymbol{u}$ is a non-negative vector satisfying $\sum_{i=1}^n u_i = 1$. Let $\mathrm{wt}(\mathcal{I}) = \sum_{i \in \mathcal{I}} u_i$ and let $\mathrm{wt}(\mathcal{O}) = \sum_{i \in \mathcal{O}} u_i$, where $\mathcal{O} \triangleq [n] \setminus \mathcal{I}$ is the set of outliers. Then, $\mathrm{wt}(\mathcal{I}) + \mathrm{wt}(\mathcal{O}) = 1$.

By contradiction, we show that if $\mathrm{wt}(\mathcal{I}) < \alpha$, then there exists a pseudo-distribution satisfying $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ that achieves a lower value of $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{\omega}]\right\|_2^2$, hence contradicting optimality of $\tilde{\mu}$. Towards this goal, we define a pseudo-distribution $\tilde{\mu}^\star$ which is supported on a single $(\boldsymbol{\omega}, \boldsymbol{x})$, the indicator vector $\mathbf{1}_\mathcal{I}$ and $\boldsymbol{x}^\circ$. Therefore, $\tilde{\mathbb{E}}_{\tilde{\mu}^\star}[\omega_i] = 1$ iff $i \in \mathcal{I}$ and zero otherwise. Clearly, $\tilde{\mu}^\star$ satisfies $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$. Therefore, any convex combination $\tilde{\mu}_\lambda = (1-\lambda)\tilde{\mu} + \lambda\tilde{\mu}^\star$ also satisfies $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$. We now show that whenever $\mathrm{wt}(\mathcal{I}) < \alpha$, then $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}_\lambda}[\boldsymbol{\omega}]\right\|_2^2 < \left\|\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{\omega}]\right\|_2^2$ for some $\lambda > 0$, thus contradicting optimality of $\tilde{\mu}$. We observe that:

$$\boldsymbol{u}_\lambda = \frac{1}{\alpha\, n}\tilde{\mathbb{E}}_{\tilde{\mu}_\lambda}[\boldsymbol{\omega}] = \frac{1}{\alpha\, n}(1-\lambda)\tilde{\mathbb{E}}_{\tilde{\mu}}[\boldsymbol{\omega}] + \frac{1}{\alpha\, n}(\lambda)\tilde{\mathbb{E}}_{\tilde{\mu}^\star}[\boldsymbol{\omega}] = (1-\lambda)\boldsymbol{u} + \frac{\lambda}{\alpha\, n}\mathbf{1}_\mathcal{I}. \qquad \text{(A183)}$$

First, we compute the squared norm of $\boldsymbol{u}_\lambda$ using (A183):

$$\|\boldsymbol{u}_\lambda\|_2^2 = \overbrace{(1-\lambda)^2 \|\boldsymbol{u}\|_2^2 + 2\lambda(1-\lambda)\frac{\mathrm{wt}(\mathcal{I})}{\alpha\,n} + \frac{\lambda^2}{\alpha\,n}}^{\text{observing } \mathbf{1}_\mathcal{I}^\mathsf{T}\mathbf{1}_\mathcal{I} = \alpha\,n \text{ and } \mathbf{1}_\mathcal{I}^\mathsf{T}\boldsymbol{u} = \mathrm{wt}(\mathcal{I})} . \tag{A184}$$

Next, we lower bound $\|\boldsymbol{u}\|_2^2$ in terms of $\mathrm{wt}(\mathcal{I})$ and $\mathrm{wt}(\mathcal{O})$. Observe that for any fixed values of $\mathrm{wt}(\mathcal{I})$ and $\mathrm{wt}(\mathcal{O})$, the minimum of $\|\boldsymbol{u}\|_2^2$ is attained by the vector $\boldsymbol{u}$ such that $u_i = \frac{1}{\alpha\,n}\mathrm{wt}(\mathcal{I})$ for each $i \in \mathcal{I}$ and $u_i = \frac{1}{(1-\alpha)n}\mathrm{wt}(\mathcal{O})$ otherwise. This gives:

$$\|\boldsymbol{u}\|_2^2 \geq \overbrace{\left(\frac{\mathrm{wt}(\mathcal{I})}{\alpha\,n}\right)^2 \alpha\,n}^{\text{sum of } u_i^2 \text{ for } i \in \mathcal{I}} + \overbrace{\left(\frac{1-\mathrm{wt}(\mathcal{I})}{(1-\alpha)n}\right)^2 (1-\alpha)n}^{\text{sum of } u_i^2 \text{ for } i \in \mathcal{O}} \tag{A185}$$

$$= \frac{\mathrm{wt}(\mathcal{I})^2}{\alpha\,n} + \frac{(1-\mathrm{wt}(\mathcal{I}))^2}{(1-\alpha)n}$$

$$= \frac{1}{\alpha\,n} \cdot \left(\mathrm{wt}(\mathcal{I})^2 + (1-\mathrm{wt}(\mathcal{I}))^2\left(\frac{\alpha}{1-\alpha}\right)\right). \tag{A186}$$

Combining (A184) and (A186):

$$\|\boldsymbol{u}_\lambda\|_2^2 - \|\boldsymbol{u}\|_2^2 = \overbrace{(-2\lambda + \lambda^2) \|\boldsymbol{u}\|_2^2}^{=(1-\lambda)^2\|\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}\|_2^2} + 2\lambda(1-\lambda)\frac{\mathrm{wt}(\mathcal{I})}{\alpha\,n} + \frac{\lambda^2}{\alpha\,n} \tag{A187}$$

$$\overset{\overbrace{\leq}^{\text{since } (-2\lambda + \lambda^2) \leq 0}}{} \frac{-2\lambda + \lambda^2}{\alpha\,n}\left(\mathrm{wt}(\mathcal{I})^2 + (1-\mathrm{wt}(\mathcal{I}))^2\frac{\alpha}{1-\alpha}\right) + 2\lambda(1-\lambda)\frac{\mathrm{wt}(\mathcal{I})}{\alpha\,n} + \frac{\lambda^2}{\alpha\,n}. \tag{A188}$$

Rearranging (note that this part slightly differs from [37], but with the same conclusion):

$$\|\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}_\lambda\|_2^2 \geq \frac{\lambda}{\alpha\,n}\left((2-\lambda)\left(\mathrm{wt}(\mathcal{I})^2 + (1-\mathrm{wt}(\mathcal{I}))^2\frac{\alpha}{1-\alpha}\right) - 2(1-\lambda)\mathrm{wt}(\mathcal{I}) - \lambda\right) \tag{A189}$$

$$= \frac{\lambda(2-\lambda)}{\alpha\,n}\left(\left(\mathrm{wt}(\mathcal{I})^2 + (1-\mathrm{wt}(\mathcal{I}))^2\frac{\alpha}{1-\alpha}\right) - \frac{2(1-\lambda)}{(2-\lambda)}\mathrm{wt}(\mathcal{I}) - \frac{\lambda}{(2-\lambda)}\right) \tag{A190}$$

$$\text{observing } \frac{2(1-\lambda)}{(2-\lambda)} = \frac{2(1-\lambda)}{2(1-\lambda) + \lambda} < 1 \text{ (for } 0 < \lambda \leq 1)$$

$$\text{and } \frac{1}{(2-\lambda)} \leq 1 < \frac{1-\mathrm{wt}(\mathcal{I})}{1-\alpha} \text{ (for } 0 \leq \mathrm{wt}(\mathcal{I}) < \alpha \leq 1)$$

$$> \frac{\lambda(2-\lambda)}{\alpha\,n}\left(\mathrm{wt}(\mathcal{I})^2 + (1-\mathrm{wt}(\mathcal{I}))^2\frac{\alpha}{1-\alpha} - \mathrm{wt}(\mathcal{I}) - \frac{1-\mathrm{wt}(\mathcal{I})}{1-\alpha}\lambda\right) \tag{A191}$$

$$= \frac{\lambda(2-\lambda)}{\alpha\,n}\left(-\mathrm{wt}(\mathcal{I})(1-\mathrm{wt}(\mathcal{I})) + (1-\mathrm{wt}(\mathcal{I}))^2\frac{\alpha}{1-\alpha} - \frac{1-\mathrm{wt}(\mathcal{I})}{1-\alpha}\lambda\right) \tag{A192}$$

$$= \frac{\lambda(2-\lambda)(1-\mathrm{wt}(\mathcal{I}))}{\alpha\,n(1-\alpha)}\left(-\mathrm{wt}(\mathcal{I})(1-\alpha) + (1-\mathrm{wt}(\mathcal{I}))\alpha - \lambda\right) \tag{A193}$$

$$= \overbrace{\frac{\lambda(2-\lambda)(1-\mathrm{wt}(\mathcal{I}))}{\alpha\,n(1-\alpha)}}^{\geq 0}\left(\alpha - \mathrm{wt}(\mathcal{I}) - \lambda\right). \tag{A194}$$

90

Now whenever $\mathrm{wt}(\mathcal{I}) < \alpha$, $(\alpha - \mathrm{wt}(\mathcal{I}) - \lambda) > 0$ for a sufficiently small λ. Thus we can choose a small enough $\lambda > 0$ such that $\|\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}_\lambda\|_2^2 > 0$, which contradicts optimality of $\tilde{\mu}$. ∎

Using Lemma A32 and Lemma A36 we can finally prove the correctness of Theorem 17. Let $\tilde{\mu}$ be a pseudo-distribution satisfying $\mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}}$ that minimizes $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\boldsymbol{\omega}\right]\right\|_2^2$. Such a pseudo-distribution exists since the set contains at least the distribution with $\omega_i = 1$ iff $i \in \mathcal{I}$ and $\boldsymbol{x} = \boldsymbol{x}^\circ$.

From Lemma A32, we have $\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \frac{\alpha\eta M_x}{2}$. Let $Z \doteq \sum_{i \in \mathcal{I}}\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{|\mathcal{I}|}$ (this is a normalization factor, such that $\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{Z|\mathcal{I}|}$ is a valid pdf over the inliers, $i.e.$, sums up to 1). By a rescaling, we obtain:

$$\sum_{i \in \mathcal{I}}\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{Z|\mathcal{I}|}\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \frac{1}{Z}\frac{\alpha\eta M_x}{2}. \tag{A195}$$

Using Lemma A36, $Z \geq \alpha$. Therefore,

$$\sum_{i \in \mathcal{I}}\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{Z|\mathcal{I}|}\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \frac{\eta M_x}{2}. \tag{A196}$$

Let $i \in [n]$ be chosen with probability $\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{\alpha\,n}$. Then, we sample $i \in \mathcal{I}$ with probability $Z \geq \alpha$. By Markov's inequality:

$$\mathbb{P}\left(\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \eta M_x\right) = \mathbb{P}\left(\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \eta M_x | i \in \mathcal{I}\right) \cdot \overbrace{\mathbb{P}\left(i \in \mathcal{I}\right)}^{\geq \alpha} \tag{A197}$$

$$\geq \alpha \cdot \mathbb{P}\left(\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \eta M_x | i \in \mathcal{I}\right) \tag{A198}$$

Markov's inequality: $\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}\left[X\right]}{a} \iff \mathbb{P}\left(X \leq a\right) \geq 1 - \frac{\mathbb{E}\left[X\right]}{a}$

$$\geq \alpha\left(1 - \frac{1}{\eta M_x}\mathbb{E}_{i \in \mathcal{I}}\left[\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2\right]\right) = \alpha\left(1 - \frac{1}{\eta M_x}\sum_{i \in \mathcal{I}}\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\omega_i\right]}{Z|\mathcal{I}|}\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2\right) \tag{A199}$$

$$\overbrace{\geq}^{\text{using (A196)}} \alpha\left(1 - \frac{1}{\eta M_x}\frac{\eta M_x}{2}\right) = \frac{\alpha}{2}. \tag{A200}$$

So we concluded that $\mathbb{P}\left(\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \eta M_x\right) \geq \frac{\alpha}{2}$ (this is the probability that a single draw satisfies $\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \eta M_x$). Calling S (as in "success") the event that $\|\boldsymbol{v}_i - \boldsymbol{x}^\circ\|_2 \leq \eta M_x$, we get that the probability of S after m draws is:

$$\mathbb{P}\left(S_m\right) = 1 - \overbrace{(1 - \mathbb{P}\left(S\right))^m}^{\text{failing } m \text{ times}} \geq 1 - \left(1 - \frac{\alpha}{2}\right)^m \tag{A201}$$

Finally, choosing the number of draws $m \geq \frac{N}{\alpha}$, we obtain

$$\mathbb{P}\left(S_m\right) \geq 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{N}{\alpha}} \tag{A202}$$

which matches the first claim in Theorem 17.

91

Now the final claim (*i.e.,* the claim that (A182) is satisfied with probability at least 0.99 for $\alpha \geq 0.01$ and $N = 10$) is just a particularization of (A202) to the given choice of N. In particular, we first observe that the probability of success $1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{N}{\alpha}}$ is a non-decreasing function of α. Then we note that the function $f(\alpha, N) \triangleq 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{N}{\alpha}}$ evaluated at $\alpha = 0.01$ and $N = 10$ is such that $f(0.01, 10) \geq 0.99$, which concludes the proof. $\blacksquare$

J   Sparse LIst-Decodable Estimation (SLIDE)

In this appendix, we present a variant of Algorithm 5 that empirically returns an accurate list of estimates, as shown in Section 8.5. We start by stating the algorithm, whose pseudocode is given in Algorithm A1.

Algorithm A1: Sparse LIst-Decodable Estimation (SLIDE).

Input: input data $(\boldsymbol{y}_i, \boldsymbol{A}_i)$, $i \in [n]$, inlier rate α.
Output: list of estimates of $\boldsymbol{x}^{\circ}$.
/* Algorithm solves a relaxation of the following problem: */
/*
$$\min_{\boldsymbol{\omega}, \boldsymbol{x}} \|\boldsymbol{\omega}\|_2^2, \quad \texttt{s.t.} \ \ \mathcal{T}_{\boldsymbol{\omega},\boldsymbol{x}} \triangleq \left\{ \begin{array}{c} \omega_i^2 = \omega_i, \ \ i = [n] \\ \sum_{i=1}^n \omega_i = \alpha n \\ \omega_i \cdot \left\|\boldsymbol{y}_i - \boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{x}\right\|_2^2 \leq \bar{c}^2, \ \ i = [n] \\ \boldsymbol{x} \in \mathbb{X} \end{array} \right\} \quad \text{(LDR)}$$

/* Compute matrix $\boldsymbol{X}^{\star}$ by solving SDP from sparse moment relaxation   */
1 $\boldsymbol{X}^{\star} = \mathsf{solve\_sparse\_moment\_relaxation\_at\_order\_2}\,(\text{LDR})$
/* Compute list of estimates   */
2 create empty list $\mathcal{L} = \emptyset$
3 **for** $i \in [n]$ **do**
4 $\quad \boldsymbol{v}_i = \left\{ \begin{array}{ll} \frac{\boldsymbol{X}^{\star}_{[\omega_i \boldsymbol{x}]}}{\boldsymbol{X}^{\star}_{[\omega_i]}} & \text{if } \boldsymbol{X}^{\star}_{[\omega_i]} > 0 \\ \boldsymbol{0} & \text{otherwise} \end{array} \right.$
5 $\quad \boldsymbol{x}_i = \mathsf{project\_to\_}\mathbb{X}(\boldsymbol{v}_i)$
6 $\quad$ add $\boldsymbol{x}_i$ to $\mathcal{L}$
7 **end**
8 **return** $\mathcal{L}$.

The proposed algorithm is very close to Algorithm 5 (and is still based on the key insights from [37]), but includes three small but important changes. First of all, instead of solving a moment relaxation of order $r = 2$, which is still expensive for large n,[43] we develop a sparse relaxation (line 1). The sparse relaxation uses the following sparse monomial basis

$$\boldsymbol{m}(\boldsymbol{\omega}, \boldsymbol{x}) \triangleq [1 \,;\, \omega_1 \,;\, \dots \,;\, \omega_n \,;\, \boldsymbol{x} \,;\, \omega_1 \boldsymbol{x} \,;\, \dots \,;\, \omega_n \boldsymbol{x}], \tag{A203}$$

---

[43]An order-2 moment relaxation of (LDR) entails solving an SDP with a matrix of size $\binom{n+d_x+2}{2}$, which is already as large as 1830 for $d_x = 9$ and $n = 50$, which is the typical setup considered in our experiments. In our tests, MOSEK [105] runs out of memory when fed an SDP of size larger than 1000.

which neglects other degree-2 monomials (*e.g.*, $\omega_i \cdot \omega_j$) that do not appear in problem (LDR) while still giving access to the pseudo-expectations used in Algorithm 5. Note that the sparse relaxation leads to SDPs of more manageable size $(n+1)(d_x+1)$.[44] Note that the idea of using a sparse moment relaxation is not new (see Remark 10 in [30]), but our relaxation is slightly different from [30, 34] and tailored to list-decodable estimation. Later in this section, we provide a derivation of the sparse moment relaxation for the rotation search problem.

The second modification is to round the estimates to the domain $\mathbb{X}$ (line 5). The latter is a consequential change: we empirically noticed that the original approach in Algorithm 5 (with our sparse relaxation) produces estimates with norm close to zero, hence leading to large estimation errors. Projecting the estimates to $\mathbb{X}$ has the effect of re-normalizing the result and correcting scaling problems, enabling the compelling results in Section 8.5. At the end of this section we show that projecting to the domain $\mathbb{X}$ is straightforward in the rotation search problem.

Finally, the third modification with respect to Algorithm 5, is that Algorithm A1 always returns n hypotheses (line 3), rather than sampling; this makes the result deterministic and independent on the choice of number of hypotheses (which can no longer be guided by the guarantees in Theorem 17).[45]

Sparse moment relaxation for rotation search. Here we provide an example of sparse relaxation arising when applying SLIDE to the rotation search problem. Let us start by tailoring the polynomial optimization problem (LDR) to rotation search:

$$\min_{\boldsymbol{\omega}, \boldsymbol{R}} \|\boldsymbol{\omega}\|_2^2, \quad \text{s.t.} \quad \left\{ \begin{array}{l} \omega_i^2 = \omega_i, \ \ i = [n] \\ \sum_{i=1}^n \omega_i = \alpha n \\ \omega_i \cdot \|\boldsymbol{b}_i - \boldsymbol{R}\boldsymbol{a}_i\|^2 \leq \bar{c}^2, \ \ i = [n] \\ \boldsymbol{R} \in \mathrm{SO}(3) \end{array} \right\} \tag{A204}$$

where we substituted the residual errors ($\left\|\boldsymbol{y}_i - \boldsymbol{A}_i^\mathsf{T}\boldsymbol{x}\right\|_2^2$) with their expression in the rotation search problem ($\|\boldsymbol{b}_i - \boldsymbol{R}\boldsymbol{a}_i\|^2$), and where we made explicit that the domain is $\mathrm{SO}(3)$. Before presenting the relaxation, we reparametrize (A204) using unit quaternions: while we could directly relax (A204) following the approach we describe below, using quaternions has the benefit of (i) leading to an even smaller relaxation (since the quaternion is parametrized by $d_x = 4$ variables instead of 9 variables needed to write a rotation matrix) and (ii) admitting a straightforward projection to the domain $\mathbb{X}$.

Proposition A37 (Quaternion-based reformulation of (A204)). *The polynomial optimization problem (A204) can be equivalently written as:*

$$\min_{\boldsymbol{\omega}, \boldsymbol{q}} \|\boldsymbol{\omega}\|_2^2, \quad s.t. \quad \left\{ \begin{array}{l} \omega_i^2 = \omega_i, \ \ i = [n] \\ \sum_{i=1}^n \omega_i = \alpha n \\ \omega_i \cdot \left(\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 - 2\mathrm{tr}\left(\boldsymbol{M}_{ij}^\mathsf{T}\boldsymbol{q}\boldsymbol{q}^\mathsf{T}\right) \right) \leq \bar{c}^2, \ \ i = [n] \\ \|\boldsymbol{q}\|^2 = 1 \end{array} \right\} \tag{A205}$$

such that any optimal solution of (A205) can be mapped back to an optimal solution of (A204) and vice-versa. In (A205), $\boldsymbol{M}_{ij}$ is a constant matrix whose expression depends on $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$.

---

[44]For instance, when $d_x = 9$ and $n = 50$, the sparse relaxation leads to a more compact SDP with a moment matrix of size 510×510.

[45]Note that we can safely discard the hypotheses corresponding to $\boldsymbol{X}_{[\omega_i]}^\star = 0$ since those are uninformative (*i.e.*, they always correspond to $\boldsymbol{v}_i = \boldsymbol{0}$). We only keep them in Algorithm A1 for the sake of simplicity, such that the output list $\mathcal{L}$ has always size n.

Proof. The proof proceeds by inspection, by reparametrizing the rotation $\boldsymbol{R}$ with the corresponding unit quaternion. Each unit quaternion corresponds to a unique rotation, hence we replace the domain $\boldsymbol{R} \in \mathrm{SO}(3)$ with the constraint that the quaternion must have unit norm (*i.e.,* $\|\boldsymbol{q}\|^2 = 1$). Comparing (A204) and (A205), we realize we only have to rewrite the maximum-residual inequality constraint in (A204) in the quaternion-based form in (A205). This derivation is largely inspired by [34] (which presents a similar reformulation applied to a different polynomial optimization problem), but here we present a simpler proof. We start by observing that the rotation matrix associated to the quaternion $\boldsymbol{q} = [q_1 \,;\, q_2 \,;\, q_3 \,;\, q_4]$ (in our notation, q_4 is the scalar part of the quaternion) is:

$$\boldsymbol{R} = \begin{bmatrix} 2(q_1^2 + q_4^2) - 1 & 2(q_1q_2 - q_3q_4) & 2(q_1q_3 + q_2q_4) \\ 2(q_1q_2 + q_3q_4) & 2(q_2^2 + q_4^2) - 1 & 2(q_2q_3 - q_1q_4) \\ 2(q_1q_3 - q_2q_4) & 2(q_2q_3 + q_1q_4) & 2(q_3^2 + q_4^2) - 1 \end{bmatrix} = \tag{A206}$$

$$\begin{bmatrix} q_1^2 + q_4^2 - q_2^2 - q_3^2 & 2(q_1q_2 - q_3q_4) & 2(q_1q_3 + q_2q_4) \\ 2(q_1q_2 + q_3q_4) & q_2^2 + q_4^2 - q_1^2 - q_3^2 & 2(q_2q_3 - q_1q_4) \\ 2(q_1q_3 - q_2q_4) & 2(q_2q_3 + q_1q_4) & q_3^2 + q_4^2 - q_1^2 - q_2^2 \end{bmatrix} \tag{A207}$$

where the expression in (A207) is obtained by substituting $\|\boldsymbol{q}\|^2 = q_1^2 + q_2^2 + q_3^2 + q_4^2 = 1$ (instead of 1) in the diagonal entries of the expression in (A206). Now, by inspection from (A207), we note that:

$$vec(\boldsymbol{R}) = \boldsymbol{P} \cdot vec(\boldsymbol{q}\boldsymbol{q}^\mathsf{T}) \tag{A208}$$

where:

$$\boldsymbol{P} \triangleq \left[\begin{array}{cccc|cccc|cccc|cccc} 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right], \quad vec(\boldsymbol{q}\boldsymbol{q}^\mathsf{T}) = \begin{bmatrix} q_1^2 \\ q_2q_1 \\ q_3q_1 \\ q_4q_1 \\ q_2q_1 \\ q_2^2 \\ q_3q_2 \\ q_4q_2 \\ q_3q_1 \\ q_3q_2 \\ q_3^2 \\ q_4q_3 \\ q_4q_1 \\ q_4q_2 \\ q_4q_3 \\ q_4^2 \end{bmatrix}$$

Equipped with these relations, we are now ready to rewrite the inequality constraint in (A204) as in (A205). We develop the squared residual $\|\boldsymbol{b}_i - \boldsymbol{R}\boldsymbol{a}_i\|^2$ in (A204) as follows:

$$\|\boldsymbol{b}_i - \boldsymbol{R}\boldsymbol{a}_i\|^2 = \quad \text{(developing the squares)} \tag{A209}$$

$$\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 - 2\boldsymbol{b}_i^\mathsf{T}\boldsymbol{R}\boldsymbol{a}_i = \quad \text{(recalling that for a scalar } a = vec(a) \text{)} \tag{A210}$$

$$\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 - 2vec(\boldsymbol{b}_i^\mathsf{T}\boldsymbol{R}\boldsymbol{a}_i) = \quad \text{(using } vec(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{C}^\mathsf{T} \otimes \boldsymbol{A})vec(\boldsymbol{B})) \tag{A211}$$

$$\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 - 2(\boldsymbol{a}_i^\mathsf{T} \otimes \boldsymbol{b}_i^\mathsf{T})vec(\boldsymbol{R}) = \quad \text{(using (A208))} \tag{A212}$$

$$\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 - 2(\boldsymbol{a}_i^\mathsf{T} \otimes \boldsymbol{b}_i^\mathsf{T})\boldsymbol{P} \, vec(\boldsymbol{q}\boldsymbol{q}^\mathsf{T}) = \quad \text{(using } \mathrm{tr}\left(\boldsymbol{A}^\mathsf{T}\boldsymbol{B}\right) = vec(\boldsymbol{A})^\mathsf{T}vec(\boldsymbol{B})) \tag{A213}$$

$$\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 - 2\mathrm{tr}\left(\boldsymbol{M}_{ij}^\mathsf{T}\boldsymbol{q}\boldsymbol{q}^\mathsf{T}\right) \tag{A214}$$

where $\boldsymbol{M}_{ij}$ is a 4×4 matrix such that $vec(\boldsymbol{M}_{ij}) = ((\boldsymbol{a}_i^\mathsf{T} \otimes \boldsymbol{b}_i^\mathsf{T})\boldsymbol{P})^\mathsf{T} = \boldsymbol{P}^\mathsf{T}(\boldsymbol{a}_i \otimes \boldsymbol{b}_i)$ (in other words, $\boldsymbol{M}_{ij}$ simply rearranges the 16 entries of the vector $\boldsymbol{P}^\mathsf{T}(\boldsymbol{a}_i \otimes \boldsymbol{b}_i)$ into a 4×4 matrix). Replacing $\|\boldsymbol{b}_i - \boldsymbol{R}\boldsymbol{a}_i\|^2$ in (A204) with (A214) yields the inequality in (A205), hence proving the claim. ∎

Proposition A38 (Sparse moment relaxation of (A205))**.** *The following SDP is a convex relaxation of the non-convex optimization problem* (A205):

$$\min_{\boldsymbol{X} \in \mathbb{S}^{5(n+1)}} \quad \sum_{i=1}^{n} \boldsymbol{X}_{[\omega_i]}^2 \tag{A215}$$

$$s.t. \quad \boldsymbol{X}_{[\omega_i, \omega_i]} = \boldsymbol{X}_{[\omega_i]}, \quad i = [n] \tag{A216}$$

$$\sum_{i=1}^{n} \boldsymbol{X}_{[\omega_i]} = \alpha n \tag{A217}$$

$$\boldsymbol{X}_{[\omega_i]} \cdot \left(\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 \right) - 2\mathrm{tr}\left(\boldsymbol{M}_{ij}^\mathsf{T} \boldsymbol{X}_{[\boldsymbol{q}, \omega_i \boldsymbol{q}^\mathsf{T}]} \right) \leq \bar{c}^2, \quad i = [n] \tag{A218}$$

$$\mathrm{tr}\left(\boldsymbol{X}_{[\boldsymbol{q}, \boldsymbol{q}^\mathsf{T}]} \right) = 1 \tag{A219}$$

$$\boldsymbol{X} \succeq 0 \tag{A220}$$

$$\boldsymbol{X}_{[1]} = 1 \tag{A221}$$

$$\boldsymbol{X}_{[\omega_i \boldsymbol{q}, \omega_i \boldsymbol{q}^\mathsf{T}]} = \boldsymbol{X}_{[\boldsymbol{q}, \omega_i \boldsymbol{q}^\mathsf{T}]}, \quad i = [n] \tag{A222}$$

$$\boldsymbol{X}_{[\omega_i \boldsymbol{q}^\mathsf{T}]} = \boldsymbol{X}_{[\omega_i, \boldsymbol{q}^\mathsf{T}]}, \quad i = [n] \tag{A223}$$

$$\boldsymbol{X}_{[\omega_i \boldsymbol{q}^\mathsf{T}]} = \boldsymbol{X}_{[\omega_i, \omega_i \boldsymbol{q}^\mathsf{T}]}, \quad i = [n] \tag{A224}$$

$$\boldsymbol{X}_{[\omega_i \boldsymbol{q}, \omega_j \boldsymbol{q}^\mathsf{T}]} = \boldsymbol{X}_{[\omega_i \boldsymbol{q}, \omega_j \boldsymbol{q}^\mathsf{T}]}^\mathsf{T}, \quad i, j = [n] \tag{A225}$$

$$\mathrm{tr}\left(\boldsymbol{X}_{[\omega_i \boldsymbol{q}, \omega_j \boldsymbol{q}^\mathsf{T}]} \right) = \boldsymbol{X}_{[\omega_i, \omega_j]}, \quad i, j = [n] \tag{A226}$$

where we index the rows of the matrix $\boldsymbol{X}$ according to the monomials $\boldsymbol{m}(\boldsymbol{\omega}, \boldsymbol{q})$, index the columns of $\boldsymbol{X}$ according to the monomials $\boldsymbol{m}(\boldsymbol{\omega}, \boldsymbol{q})^\mathsf{T} \triangleq [1, \omega_1, \dots, \omega_n, \boldsymbol{q}^\mathsf{T}, \omega_1 \boldsymbol{q}^\mathsf{T}, \dots, \omega_n \boldsymbol{q}^\mathsf{T}]$, and use the notation $\boldsymbol{X}_{[i,j]}$ to access entries of the matrix with row indexed by monomial i and column indexed by monomial j; we also overload the notation and write as $\boldsymbol{X}_{[i]}$ to denote $\boldsymbol{X}_{[i,1]}$.

Proof. While the SDP appears to be quite complicated, its constraints should become apparent from the structure of the moment matrix built on the sparse monomial basis $\boldsymbol{m}(\boldsymbol{\omega}, \boldsymbol{q}) \triangleq [1; \omega_1; \dots; \omega_n; \boldsymbol{q}; \omega_1 \boldsymbol{q}; \dots; \omega_n \boldsymbol{q}]$:

$$\boldsymbol{X} = \boldsymbol{m}(\boldsymbol{\omega}, \boldsymbol{q}) \boldsymbol{m}(\boldsymbol{\omega}, \boldsymbol{q})^\mathsf{T} = \begin{array}{c} \\ \begin{array}{c} 1 \\ \omega_1 \\ \vdots \\ \omega_n \\ \boldsymbol{q} \\ \omega_1 \boldsymbol{q} \\ \vdots \\ \omega_n \boldsymbol{q} \end{array} \begin{array}{ccccccccc} 1 & \omega_1 & \dots & \omega_n & \boldsymbol{q}^\mathsf{T} & \omega_1 \boldsymbol{q}^\mathsf{T} & \dots & \omega_n \boldsymbol{q}^\mathsf{T} \\ \left[\begin{array}{cccc|ccc} 1 & \omega_1 & \dots & \omega_n & \boldsymbol{q}^\mathsf{T} & \omega_1 \boldsymbol{q}^\mathsf{T} & \dots & \omega_n \boldsymbol{q}^\mathsf{T} \\ * & \omega_1^2 & \dots & \vdots & \omega_1 \boldsymbol{q}^\mathsf{T} & \omega_1^2 \boldsymbol{q}^\mathsf{T} & \dots & \omega_1 \omega_n \boldsymbol{q}^\mathsf{T} \\ * & * & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ * & * & \dots & \omega_n^2 & \omega_n \boldsymbol{q}^\mathsf{T} & \omega_1 \omega_n \boldsymbol{q}^\mathsf{T} & \dots & \omega_n^2 \boldsymbol{q}^\mathsf{T} \\ \hline * & * & \dots & * & \boldsymbol{q} \boldsymbol{q}^\mathsf{T} & \omega_1 \boldsymbol{q} \boldsymbol{q}^\mathsf{T} & \dots & \omega_n \boldsymbol{q} \boldsymbol{q}^\mathsf{T} \\ * & * & \dots & * & * & \omega_1^2 \boldsymbol{q} \boldsymbol{q}^\mathsf{T} & \dots & \omega_1 \omega_n \boldsymbol{q} \boldsymbol{q}^\mathsf{T} \\ * & * & \dots & * & * & * & \ddots & \vdots \\ * & * & \dots & * & * & * & \dots & \omega_n^2 \boldsymbol{q} \boldsymbol{q}^\mathsf{T} \end{array} \right] \end{array} \end{array} \tag{A227}$$

where we also reported in gray the row and column indices described in the statement of the proposition. We prove the proposition in two steps. First, we show how to rewrite (A205) using the moment matrix $\boldsymbol{X}$ in (A227), which leads to the objective and constraints in (A215)-(A219). Second, we show that any moment matrix with the structure in (A227) satisfies the constraints (A220)-(A226), hence the feasible set of (A215) contains the feasible set of (A205). Let us start by rewriting (A205) using the moment matrix $\boldsymbol{X}$:

$$\min_{\boldsymbol{\omega},\boldsymbol{q},\boldsymbol{X}} \quad \sum_{i=1}^{n} \boldsymbol{X}_{[\omega_i]}^2 \tag{A228}$$

$$\text{s.t.} \quad \boldsymbol{X}_{[\omega_i,\omega_i]} = \boldsymbol{X}_{[\omega_i]}, \quad i = [n] \tag{A229}$$

$$\sum_{i=1}^{n} \boldsymbol{X}_{[\omega_i]} = \alpha n \tag{A230}$$

$$\boldsymbol{X}_{[\omega_i]} \cdot \left(\|\boldsymbol{b}_i\|^2 + \|\boldsymbol{a}_i\|^2 \right) - 2\text{tr}\left(\boldsymbol{M}_{ij}^\mathsf{T} \boldsymbol{X}_{[\boldsymbol{q},\omega_i\boldsymbol{q}^\mathsf{T}]} \right) \leq \bar{c}^2, \quad i = [n] \tag{A231}$$

$$\text{tr}\left(\boldsymbol{X}_{[\boldsymbol{q},\boldsymbol{q}^\mathsf{T}]} \right) = 1 \tag{A232}$$

$$\boldsymbol{X} = \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q}) \cdot \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q})^\mathsf{T} \tag{A233}$$

where $\boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q}) \triangleq [1\,;\,\omega_1\,;\,\dots\,;\,\omega_n\,;\,\boldsymbol{q}\,;\,\omega_1\boldsymbol{q}\,;\,\dots\,;\,\omega_n\boldsymbol{q}]$, and we simply noticed (from inspection of (A227)) that $\boldsymbol{X}_{[\omega_i,\omega_i]} = \omega_i^2$, $\boldsymbol{X}_{[\omega_i]} = \omega_i$, $\boldsymbol{X}_{[\boldsymbol{q},\omega_i\boldsymbol{q}^\mathsf{T}]} = \omega_i \boldsymbol{q}\boldsymbol{q}^\mathsf{T}$, and $\text{tr}\left(\boldsymbol{X}_{[\boldsymbol{q},\boldsymbol{q}^\mathsf{T}]} \right) = \text{tr}\left(\boldsymbol{q}\boldsymbol{q}^\mathsf{T} \right) = \boldsymbol{q}^\mathsf{T}\boldsymbol{q} = \|\boldsymbol{q}\|^2$, hence (A228) just rewrites objective and constraints in (A205) using the entries of the moment matrix $\boldsymbol{X}$ in (A203). Problem (A228) is equivalent to (A205) and is still non-convex due to the non-convexity of the constraint $\boldsymbol{X} = \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q}) \cdot \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q})^\mathsf{T}$.

Now we are only left to prove that the feasible set of (A215) contains the feasible set of (A228). More precisely, we prove that any matrix that satisfies $\boldsymbol{X} = \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q}) \cdot \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q})^\mathsf{T}$ also satisfies constraints (A220)-(A226) in (A215). Clearly, any $\boldsymbol{X} = \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q}) \cdot \boldsymbol{m}(\boldsymbol{\omega},\boldsymbol{q})^\mathsf{T}$ is such that $\boldsymbol{X} \succeq 0$. The rest of the constraints can be also seen to hold by simple inspection of the entries of the moment matrix (A203) and recalling that our constraint set also imposes $\omega_i^2 = \omega_i$ (for all $i \in [n]$) and $\text{tr}\left(\boldsymbol{q}\boldsymbol{q}^\mathsf{T} \right) = \|\boldsymbol{q}\|^2 = 1$. Therefore, since (A215) has the same objective of (A228), but its feasible set includes the feasible set of (A228), problem (A215) is a relaxation of (A228). Finally, we observe that (A215) is a convex program, since it minimizes a convex cost function over the cone of positive-semidefinite matrices and subject to a set of linear constraints. ∎

Rounding for rotation search. According to Algorithm A1, after solving the sparse moment relaxation and obtaining the matrix $\boldsymbol{X}^\star$, we build the vectors $\boldsymbol{v}_i$ from the entries of the matrix $\boldsymbol{X}^\star$, and then project those vectors to the domain $\mathbb{X}$. In our quaternion-based formulation of the rotation search problem (Proposition A37), $\boldsymbol{v}_i$ are 4-dimensional vectors, while $\mathbb{X}$ is the set of unit quaternions. Hence projecting onto the domain $\mathbb{X}$ (line 5 in Algorithm A1) only requires normalizing the vectors $\boldsymbol{v}_i$ to have unit norm, *i.e.*, $\boldsymbol{x}_i = \boldsymbol{v}_i/\|\boldsymbol{v}_i\|$. In particular, we add $\boldsymbol{x}_i = \boldsymbol{v}_i/\|\boldsymbol{v}_i\|$ whenever $\|\boldsymbol{v}_i\| > 0$, while we mark an estimate as invalid when $\|\boldsymbol{v}_i\| = 0$ and disregard it from the evaluation.