arXiv:2403.09465v1 [cs.DS] 14 Mar 2024

Outlier Robust Multivariate Polynomial Regression

Vipul Arora[*1] Arnab Bhattacharyya[†1] Mathews Boban[‡2] Venkatesan Guruswami[§3]
Esty Kelman[¶4]

[1]National University of Singapore. {vipul,arnab}@comp.nus.edu.sg.
[2]National University of Singapore. mathewsboban242@gmail.com
[3]University of California, Berkeley. venkatg@berkeley.edu.
[4]Massachusetts Institute of Technology, and Boston University. ekelman@mit.edu.

February 2024

Abstract

We study the problem of *robust multivariate polynomial regression*: let $p\colon \mathbb{R}^n \to \mathbb{R}$ be an unknown n-variate polynomial of degree at most d in each variable. We are given as input a set of random samples $(\mathbf{x}_i, y_i) \in [-1,1]^n \times \mathbb{R}$ that are noisy versions of $(\mathbf{x}_i, p(\mathbf{x}_i))$. More precisely, each $\mathbf{x}_i$ is sampled independently from some distribution χ on $[-1,1]^n$, and for each i independently, y_i is arbitrary (i.e., an outlier) with probability at most $\rho < 1/2$, and otherwise satisfies $|y_i - p(\mathbf{x}_i)| \leq \sigma$. The goal is to output a polynomial $\widehat{p}$, of degree at most d in each variable, within an ℓ_∞-distance of at most $O(\sigma)$ from p.

Kane, Karmalkar, and Price [FOCS'17] solved this problem for $n = 1$. We generalize their results to the n-variate setting, showing an algorithm that achieves a sample complexity of $O_n(d^n \log d)$, where the hidden constant depends on n, if χ is the n-dimensional Chebyshev distribution. The sample complexity is $O_n(d^{2n} \log d)$, if the samples are drawn from the uniform distribution instead. The approximation error is guaranteed to be at most $O(\sigma)$, and the run-time depends on $\log(1/\sigma)$. In the setting where each $\mathbf{x}_i$ and y_i are known up to N bits of precision, the run-time's dependence on N is linear. We also show that our sample complexities are optimal in terms of d^n. Furthermore, we show that it is possible to have the run-time be independent of $1/\sigma$, at the cost of a higher sample complexity.

[*]Supported in part by NRF-AI Fellowship R-252-100-B13-281.
[†]Supported in part by NRF-AI Fellowship R-252-100-B13-281, Amazon Faculty Research Award, and Google South & Southeast Asia Research Award.
[‡]Supported in part by NRF-AI Fellowship R-252-100-B13-281.
[§]Supported in part by NSF CCF-2211972 and a Simons Investigator Award
[¶]Supported in part by an Amazon Faculty Research Award to AB, in part by ERC grant 834735, and in part by NSF TRIPODS program (award DMS-2022448)

1 Introduction

"Curve fitting" or *polynomial regression* is one of the oldest and most fundamental learning problems: find a polynomial that approximately satisfies the input-output relationship displayed by a collection of data points. Polynomial regression has a vast range of applications, from the physical sciences to statistics and machine learning; see, e.g., the books [Wol06, Zie11] for discussions and references.

The focus of this work is on *multivariate* polynomial regression, which is the task of learning the class of bounded degree polynomials from random noisy samples. Multivariate polynomial regression is a natural requirement in many applications. For example, in computer vision, boundaries of objects are often modeled as low-degree bivariate polynomials, so it is well-motivated to fit curves to estimates of object boundaries. Our goal is to design *robust* regression algorithms, which can withstand having a constant fraction of the input data be arbitrary outliers in the same setting as in [AK03, GZ16, KKP17].

We next formally state the problem of robust multivariate regression. Let us denote by $\mathcal{P}_d$ the class of all n-variate *individual* degree-d polynomials, which are the polynomials with degree at most d in each variable[1].

Robust Multivariate Polynomial Regression Problem. *Let $\sigma > 0$ be a noise bound, $C > 1$ be an approximation factor, $\rho \in [0, 1]$ be the outlier probability, χ be a probability distribution over $[-1, 1]^n$. Fix an unknown $p \in \mathcal{P}_d$ and let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M)\}$ be a set random samples where for each $i \in [M]$ independently, $\mathbf{x}_i$ sampled from χ, and $y_i \in \mathbb{R}$ is an* inlier *satisfying $|y_i - p(\mathbf{x}_i)| \leq \sigma$ with probability $1 - \rho$ and otherwise, it may be an* outlier*, i.e., the noise may be arbitrarily large. The goal is to design an efficient algorithm that, given the set S of random samples as input, recovers a polynomial $\widehat{p} \in \mathcal{P}_d$ satisfying*

$$\max_{\mathbf{x} \in [-1,1]^n} |p(\mathbf{x}) - \widehat{p}(\mathbf{x})| \leq C\sigma,$$

with probability at least $1 - \delta$.

Note that though the locations of the outliers are random, i.e., each sample is an outlier with probability ρ independently, the noise for both the inliers and the outliers is still allowed to be chosen in an adversarial way (meaning an adversary can choose the values of all the y_i's after seeing the entire sample set $\{\mathbf{x}_i\}$).

In the univariate setting, recovery for non-trivial values of ρ was first shown by Guruswami and Zuckerman [GZ16] for $\rho < 1/\log d$. Previously, Arora and Khot [AK03] had shown $\rho < 1/2$ was information-theoretically necessary for unique recovery. Subsequently, Kane, Karmalkar, and Price [KKP17] designed a simple and optimal (up to constants) algorithm that runs in polynomial time for any $\rho < 1/2$, uses $\Theta(d \log d)$ samples from the Chebyshev measure on $[-1, 1]$, or $\Theta(d^2)$ uniform samples and outputs a degree-d univariate polynomial $\widehat{p}$ satisfying $\max_{x \in [-1,1]} |p(x) - \widehat{p}(x)| \leq C\sigma$. They show how to achieve C as close to 2 as desired. In addition, they show that to solve the problem for $d = 2$ with probability at least $2/3$, $C > 1.09$ is needed, while for general d, to succeed with constant probability, one needs $C > 1 + \Omega(1/d^3)$.

1.1 Main results

We wish to minimize the sample complexity M. Our algorithmic results are mainly when the measure χ is either the uniform distribution or the n-dimensional Chebyshev measure, i.e., the n-fold product of the Chebyshev measure on $[-1, 1]$, with the probability density function $\propto 1/\sqrt{1 - x^2}$ for $x \in [-1, 1]$.

[1] This is in contrast to the usual convention of the *total* degree being at most d. Note that the class of polynomials of *total* degree at most d is strictly included in $\mathcal{P}_d$. Our results (for $\mathcal{P}_d$) can be translated for the class of total degree-d polynomials; See discussion in Remark 1.7.

Note that when n is large, for some distributions, solving the multivariate polynomial regression problem requires $\exp(n)$ many samples, even for polynomials of *total* degree $d = 1$ (for completeness, we provide a proof in Appendix A). So, for sample-efficient algorithms, it is prudent to assume $n > 1$ being a constant. In this setup, then, the total degree of an individual degree-d polynomial is at most nd, i.e., $O(d)$, and hence the multivariate polynomial regression problem becomes oblivious to the degree being total or individual. Thus, we focus on learning the class $\mathcal{P}_d$ of *individual* degree-d polynomials in a constant number of variables.

We now state our main results. Denote the cube $[-1, 1]^n$ by $\mathcal{C}_n$; we will omit the subscript when the dimension is clear from the context. Let $\|\cdot\|_{\mathcal{C},\infty}$ denote the ℓ_∞ norm over $\mathcal{C}_n$.

Theorem 1.1. *Let $\sigma \geq 0, \eta > 0$, and ρ be any constant $< 1/2$. There is an algorithm that almost solves the Robust Multivariate Polynomial Regression Problem with a constant approximation factor, up to an additive error of η. The output of the algorithm is $\widehat{p} \in \mathcal{P}_d$ that satisfies*

$$|p(\mathbf{x}) - \widehat{p}(\mathbf{x})| \leq O(\sigma) + \eta, \qquad \text{for all } \mathbf{x} \in \mathcal{C}_n,$$

with probability at least $2/3$. It uses $M = O_n(d^n \log d)$ samples drawn from the multidimensional Chebyshev distribution, or $M = \widetilde{O}_n(d^{2n})$ if the samples are drawn from the uniform measure. Its run-time is at most $\mathsf{poly}(\log \|p\|_{\mathcal{C},\infty}, M, \log(1/\eta))$.

The notations $\widetilde{O}$, $\widetilde{\Theta}$ hide factors proportional to $\log d$ above; the dependence on η or σ is kept explicit.

In case σ is known to be at least 2^{-N}, one may choose $\eta = 2^{-N}$ to guarantee $\|\widehat{p} - p\|_{\mathcal{C}_n,\infty} \leq O(\sigma)$ and run-time proportional to $\mathsf{poly}(N)$. Generalizing this observation, we consider the *N-bit precision setting*, where both the sample locations $\mathbf{x}_i$ and the labels y_i are truncated to N bits of precision; this is consistent with a computational model where real numbers can only be specified up to N bits of precision. We show that in the N-bit precision setting, a variant of our algorithm achieves a constant approximation factor without any additional additive error.

Theorem 1.2. *Let N be the number of bits of precision, $\sigma \geq 2^{-N}$, and ρ be any constant $< 1/2$. There exists an algorithm for the Robust Multivariate Polynomial Regression Problem, wherein each $\mathbf{x}_i$ is now drawn from a continuous distribution χ and then rounded to N bits of precision, and each y_i is similarly rounded. The output of the algorithm is $\widehat{p} \in \mathcal{P}_d$, that satisfies*

$$|p(\mathbf{x}) - \widehat{p}(\mathbf{x})| \leq O(\sigma), \qquad \text{for all } \mathbf{x} \in \mathcal{C}_n,$$

with probability at least $2/3$. It uses $M = O_n(d^n \log d)$ samples drawn from the multidimensional Chebyshev distribution, or $M = \widetilde{O}_n(d^{2n})$ if the samples are drawn from the uniform measure. Its run-time is at most $\mathsf{poly}(\log \|p\|_{\mathcal{C},\infty}, M, N)$.

To avoid a run-time dependent on $\|p\|_{\mathcal{C},\infty}$ and $1/\eta$, in case they are unknown or too large, we also obtain a variant of the algorithm that achieves an explicit constant multiplicative approximation factor, as close to 2 as desired and independent of $\|p\|_{\mathcal{C},\infty}$ and $1/\eta$, at the cost of a higher sample complexity.

Theorem 1.3. *Let $\varepsilon > 0$, $\sigma \geq 0$, and a constant $\rho < 1/2$. There exists an algorithm that solves the Robust Multivariate Polynomial Regression Problem. The output of the algorithm is $\widehat{p} \in \mathcal{P}_d$, that satisfies*

$$|p(\mathbf{x}) - \widehat{p}(\mathbf{x})| \leq (2 + \varepsilon)\sigma \qquad \text{for all } \mathbf{x} \in \mathcal{C}_n,$$

with probability at least $2/3$. It uses $M = \mathsf{poly}(d^{n^2}, 1/\varepsilon^n)$ samples drawn from either the multidimensional Chebyshev distribution or the uniform distribution. Its run-time is $\mathsf{poly}(M)$.

We complement the above results by showing lower bounds on the sample complexity of robust multivariate polynomial regression.

Theorem 1.4. *For any approximation factor $C > 1$, there exists $c = c(C) > 0$ such that any algorithm, that solves the Robust Multivariate Polynomial Regression Problem, requires at least $(cd)^{2n}$ samples drawn from the uniform measure to succeed with probability more than $2/3$. This holds for any outlier probability ρ.*

The lower bound matches the upper bound of Theorem 1.2 up to lower order terms (in the case of uniform sampling) for constant n, and C, and holds even for $\rho = 0$, where there are no outliers. The following result shows that our result in Theorem 1.2 for the multidimensional Chebyshev measure matches the optimal sample complexity over arbitrary distributions[2].

Theorem 1.5. *For any approximation factor $C > 1$, and any outlier probability $\rho > 0$, there exists $c = c(C, \rho) > 0$ such that any algorithm, that solves the Robust Multivariate Polynomial Regression Problem, requires at least $(cd)^n \log d$ samples drawn from any measure over $[-1, 1]^n$ to succeed with probability more than $2/3$.*

A comparison of the results across parameter regimes may be given via the following table, wherein M is the sample complexity, and $C_p = \log \|p\|_{\mathcal{C},\infty}$:

Setting	Approximation	Chebyshev Measure	Uniform Measure	Run-time
Exact	$O(\sigma) + \eta$	$O_n(d^n \log d)$	$\widetilde{O}_n(d^{2n})$	$\mathsf{poly}(C_p, M, \log(1/\eta))$
N-bit	$O(\sigma)$	$\Theta_n(d^n \log d)$	$\widetilde{\Theta}_n(d^{2n})$	$\mathsf{poly}(C_p, M, N)$
Small ε	$(2 + \varepsilon)\sigma$	$\mathsf{poly}(d^{n^2}, 1/\varepsilon^n)$	$\mathsf{poly}(d^{n^2}, 1/\varepsilon^n)$	$\mathsf{poly}(M)$

Table 1: The upper bounds in the first line follow from Theorem 1.1, the second line from Theorem 1.2, and the third line from Theorem 1.3. The lower bounds in the second line for the Uniform Measure follow from Theorem 1.4, and for the Chebyshev Measure from Theorem 1.5.

Remark 1.6. *[KKP17] were able to achieve both optimal sample complexity and efficient run-time independent of $\|p\|_\infty/\sigma$ with a single algorithm in the univariate setting. In contrast, as we elaborate in the proof overview, our Theorem 1.3 incurs an additional blowup in the sample complexity; we leave open the problem of realizing the error guarantees of Theorem 1.3 with the optimal number of samples.*

Remark 1.7. *Both our upper and lower bounds hold for the class of total degree-d polynomials, when n is bounded, since total degree being at most d implies individual degree being at most d, and individual degree being at most d implies total degree being at most dn.*

1.2 Main technical contributions

Our main technical contributions are twofold, and they may be of interest more broadly. First, let $\{\mathcal{C}_j\}_{j \in [m]^n}$ be a partition of the cube $\mathcal{C}_n$ induced by the m-Chebyshev extremas on each axis. We call it the (m, n)-Chebyshev partition[3] of $\mathcal{C}_n$. For $n = 1$, [KKP17] showed how to approximate a univariate polynomial of degree at most d on $[-1, 1]$ by an appropriate piece-wise constant function with respect to the Chebyshev partition. We extend their result to the multivariate case.

Theorem 1.8. *[Approximation by piece-wise constant functions] Let $p : \mathcal{C}_n \to \mathbb{R}$ be a polynomial of degree at most d in each variable, and $m \geq d$. Let $r : \mathcal{C}_n \to \mathbb{R}$ be a piece-wise constant function with respect*

[2] Similarly, the lower bounds match the respective sample complexities of Theorem 1.1, when the additive error η approaches 0, since the algorithm's run-time grows as $\eta \to 0$, but the sample complexities remain unchanged.

[3] See formal Definition 2.3, and Figure 1 for an illustration of a 2-dimensional Chebyshev partition.

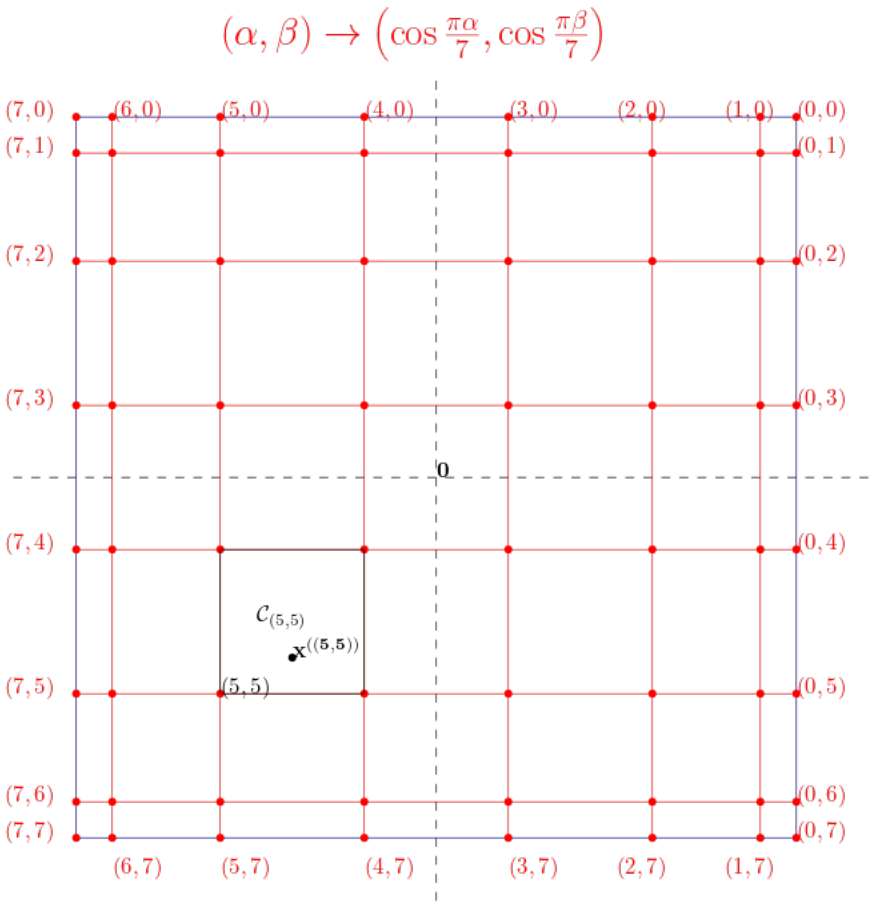

Figure 1: An illustration of a 2-dimensional $(7, 2)$-Chebyshev partition (in red) super-imposed on the 2-dimensional solid cube $\mathcal{C}_2 = [-1, 1]^2$, with boundary in blue. The cells are indexed by their bottom-left Chebyshev extremas (the red points). Theorem 1.8 essentially proves that on any cell, for example, $\mathcal{C}_{(5,5)}$ (in black), p can be well approximated by its evaluation on one arbitrary point $\mathbf{x}^{((5,5))} \in \mathcal{C}_{(5,5)}$.

to the (m, n)-Chebyshev partition, such that for every $\boldsymbol{j} \in [m]^n$, there exist a point $\mathbf{x}^{(j)} \in \mathcal{C}_{\boldsymbol{j}}$, such that $r(\mathbf{x}) = p(\mathbf{x}^{(j)})$, for all $\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}$. Then, there exists a universal constant C such that,

$$\|p - r\|_{\mathcal{C}_n,\infty} \leq C\frac{dn}{m}\|p\|_{\mathcal{C}_n,\infty}.$$

This is proved in Section 3.1. Second, we show how to relate the maximum value of a bounded degree polynomial on $\mathcal{C}_n$ with its ℓ_1 norm, on the same cube $\mathcal{C}_n$.

Theorem 1.9. *There exists a global constant $C > 0$ such that, for every polynomial $p \in \mathcal{P}_d$:*

$$\|p\|_{\mathcal{C}_n,\infty} \leq C^n d^{2n}\|p\|_{\mathcal{C}_n,1}.$$

We also prove the tightness of the relation between ℓ_∞ and ℓ_1 norms.

Proposition 1.10. *There exists a global constant $c > 0$ such that for every odd d, there exists a family of individual degree-d polynomials $\{f_n\}_{n\in\mathbb{N}}$, where $f_n : \mathcal{C}_n \to \mathbb{R}$, such that $\|f_n\|_{\mathcal{C}_n,\infty} \geq c^n d^{2n}\|f_n\|_{\mathcal{C}_n,1}$.*

These are proved in Section 5. In Appendix D, we show how to prove a similar result, using a previous work by [Wil74] and a simpler argument, but it has factors that are worse by a multiple of $\sim n^{2.5n}$.

1.3 Related Work

Given the fundamental nature of the polynomial regression problem, there is a long history of work on the problem, but mostly in the univariate setting. Arora and Khot [AK03] were the first to study this problem in our random outlier noise model, giving an algorithm that in $O(\frac{d^2}{\sigma}\log\frac{d}{\sigma})$ random noisy samples outputs an $O(\sigma)$-approximation (in ℓ_∞) to the (actual) hidden polynomial, where the outlier rate $\rho = 0$. This was improved in a work by Guruswami and Zuckerman [GZ16], who gave a computationally efficient algorithm for all $\rho < 1/\log d$. Finally, in a significant improvement, Kane, Karmalkar and Price [KKP17] obtained computationally efficient algorithms for any $\rho < 1/2$, while having no additional requirements for σ or $\|p\|_\infty$. As far as we know, Daltrophe, Dolev and Lotker [DDL18] were the first to consider the *multivariate* setting of this problem. For the two-dimensional case ($n = 2$), they gave an algorithm that with $O(\frac{d^4}{\sigma}\log\frac{d}{\sigma})$ random noisy samples outputs a $c(2)\cdot\sigma$-approximation (in ℓ_∞), for any $\rho < \frac{1}{2}$, where $c(2) = 3$. A limitation of their result is that $c(n)$ grows exponentially in n. In contrast, we obtain a constant factor approximation for all n.

There has also been a surge of recent research in the related area of robust statistics. Here, instead of the outliers being randomly placed, their locations are chosen adversarially. For the setting when the *total* degree is fixed, and the dimension n is growing, Klivans, Kothari and Meka [KKM18] gave an algorithm using the sum-of-squares method. However, their sample complexity is $\mathrm{poly}(n^d)$, which is exponential in the degree; moreover, the output guarantee is with respect to the $\|\cdot\|_2$ norm, instead of the $\|\cdot\|_\infty$ norm in our setting. Other related works in this spirit are that of Diakonikolas, Kamath, Kane, Li, Steinhardt and Stewart [DKK+19] and Prasad, Suggala, Balakrishnan and Ravikumar [PSBR20]. The work of Diakonikolas, Kong, and Stewart [DKS19] also studied the related problem of adversarially robust linear regression, but with the assumption that the $\mathbf{x}_i$'s are drawn from a Gaussian.

1.4 Technical Overview

We first sketch the algorithm designed by Kane, Karmalkar, and Price [KKP17], henceforth KKP, and their analysis for the univariate case, $n = 1$. For univariate polynomial interpolation, the points at which the noisy samples are located play an important role in determining the interpolation error. Choosing the points to be the Chebyshev nodes, which are the roots of Chebyshev polynomials (see Definition 7.1), is a good starting point, as suggested by approximation theory literature. However, the algorithm receives random samples, which may not necessarily be located at the Chebyshev nodes. Instead, KKP argue that they have enough inliers around each Chebyshev node. For this, they define a partition of the interval $[-1, 1]$ on the extremal points of Chebyshev polynomials, which they call *the size-m Chebyshev partition*. In their algorithm and its analysis, they assume that the set of samples is *good*, in the sense that in every part of the partition, there is only a small fraction of outliers; this good event is guaranteed to happen with high probability.

1.4.1 KKP's Algorithm and Analysis

Formally, the size-m Chebyshev partition of $[-1, 1]$ is the set of intervals $I_j = \left[\cos\frac{\pi j}{m}, \cos\frac{\pi(j-1)}{m}\right]$, for all $j \in [m]$. Given a set of s samples (x_i, y_i) where x_i's are drawn from some distribution over $[-1, 1]$, and y_i's are the corresponding labels, the algorithm uses the idea of *median-based recovery*. For every interval I_j:

- Let $\tilde{y}_j$ be the median of y_i's of samples for which $x_i \in I_j$. Since the set of samples is assumed to be *good*, i.e., the fraction of outliers in each interval is strictly less than $1/2$, $\tilde{y}_j$ lies in between two inliers located in the interval I_j.
- Let $\tilde{x}_j$ be an arbitrary point in I_j.

- Let $\widehat{p}$ be a minimizer, over all degree-d polynomials, of the empirical ℓ_∞ error $\max_j |\widehat{p}(\tilde{x}_j) - \tilde{y}_j|$ over all $j \in [m]$.

As m grows, the partition gets finer, and the error gets better, though at a cost of higher sample complexity. Iteratively applying the median-based recovery on the residual left from previous iteration improves the approximation, and in $\log\left(\max_{x\in[-1,1]} |p(x)|/\sigma\right)$ iterations, the error drops down to 3σ.

The backbone of their analysis is a technical result for approximating p on a size-m Chebyshev partition, by a piece-wise constant function (with respect to the same partition) that matches p on at least one point in every part of the partition.

Lemma 1.11. *[Lemma 2.1, [KKP17]] Let $g : \mathbb{R} \to \mathbb{R}$ be a (univariate) degree-d polynomial. Let $\{I_j\}_{j\in[m]}$ denote the m-Chebyshev partition of $[-1, 1]$, for some $m \geq d$. Let $r : [-1, 1] \to \mathbb{R}$ be piece-wise constant, so that for each $k \in [m]$, there exists $x^*_k \in I_k$, such that $r(x) = g(x^*_k)$ for all $x \in I_k$. Then, there exists a universal constant C such that, for any $q \geq 1$,*

$$\|g - r\|_q \leq \frac{Cd}{m}\|g\|_q.$$

To prove Lemma 1.11, they used Nevai's inequality [Nev79], an ℓ_q-version of Bernstein's inequality, to bound the ℓ_q approximation error by a multiple of the ℓ_q norm of p. The multiple is linear in the degree d, and $1/m$. The bound from Nevai's inequality works for all "inner" parts of the Chebyshev partition, as it relies on the fact that the length of any part I_j, where $j \notin \{1, m\}$, is at most $O(\sqrt{1-x^2}/m)$, for every $x \in I_j$. To bound the approximation error on the peripheral parts I_1, I_m, they use Markov Brothers' inequality (Lemma 2.2). Here they strongly rely on the fact that those parts are much narrower[4] ($|I_1| = |I_m| = O(1/m^2)$) than the inner parts. This additional $1/m$ factor in the length compensates for the worse bound from Lemma 2.2.

Lemma 1.11, with q set to ∞, is used to bound the error of the median-based recovery procedure, in terms of $\|p\|_{\mathcal{C}_1,\infty}$. This allows the $\log \frac{\|p\|_{\mathcal{C}_1,\infty}}{\sigma}$ iterations to be all that is further needed to bring the error down to 3σ.

To avoid the run-time dependence on $\max_{x\in[-1,1]} |p(x)|/\sigma$, which maybe unknown or too large, KKP first run an ℓ_1 regression, which gives an ℓ_∞ error of at most $O(d^2\sigma)$, and then run the median-based recovery algorithm on the residual polynomial, which in $\log d$ iterations drops the error further to at most 3σ. Lemma 1.11, with $q = 1$, is used to bound the ℓ_1-error of the ℓ_1-minimizer by $O(\sigma)$. A further application of Lemma 2.2 bounds the ℓ_∞-error of the ℓ_1-minimizer by $O(d^2\sigma)$. This then allows for a bound of $\log d$ on the number of iterations needed, and hence the algorithm's run-time.

1.4.2 Our Results

Generalizing to the multivariate case ($n > 1$): We show that the idea of KKP generalizes to the multivariate setting by considering a tensorization of the Chebyshev partition, i.e., we divide the cube $[-1, 1]^n$ into m^n cells according to a grid partition, where each axis is divided into m intervals, according to the size-m Chebyshev partition of $[-1, 1]$ defined by KKP. The analysis takes steps similar to the analysis done by KKP, and some of the proofs follow by 'tensoring' KKP's arguments in some sense.

There are some subtleties that we take care of along the way. We successfully show optimal sample complexity results, in terms of the dependence on d^n, for the median-based recovery algorithm, while for the ℓ_1 regression, we need more samples. For this reason, we first analyze the median based recovery algorithm, the running time (but not the sample complexity) of which, depends on $\max_{\boldsymbol{x}\in\mathcal{C}_n} |p(\boldsymbol{x})|$. Later, we show that

[4] A pictorial demonstration of this narrowness, for 2-dimensional partitions, can be observed in Figure 1.

by running the ℓ_1 regression on weighted averages (with respect to cells) as the first step, we reach a constant approximation factor in bounded run-time at the cost of increasing the exponent in the sample complexity from n to $O(n^2)$.

Overview of the algorithms and analyses: For using the median-recovery algorithm, we devise a multivariate analog (Theorem 1.8) of Lemma 1.11 for the ℓ_∞ norm. Specifically, we show that, for large enough m, every n-variate, individual degree-d polynomial p is well approximated by any piece-wise constant function with respect to the (m, n)-Chebyshev partition that matches p on at least one point in each cell. This is proved by a repeated application of the *univariate* ℓ_∞ approximation statement from Lemma 1.11. Algorithmically, we then do median-based recovery on a fine enough Chebyshev partition of $\mathcal{C}_n$, and iteratively improve the output of the ℓ_∞ regression. After at most $\log(\|p\|_{\mathcal{C}_n,\infty}/\eta)$ iterations, we achieve an $O(\sigma) + \eta$ approximation. A $\mathrm{poly}(\log \|p\|_{\mathcal{C},\infty}, M, \log(1/\eta))$ run-time is thus achieved. One may set $\eta = \sigma$ to achieve an $O(\sigma)$ approximation, in this case the run-time is dependent on $\log(1/\sigma)$ instead.

We also consider the *finite bit precision* setting where the samples are represented using at most N bits of precision. This forces $\sigma \geq 2^{-N}$, and the (location, evaluation) pairs of the random input samples are now rounded to N bits. In this case, the samples' locations are not exact, and hence we are uncertain as to which Chebyshev cell they belong to. To deal with it, we discard samples that lie in a small ℓ_1 neighborhood of the boundary of the cells. (See Figure 2 for an illustration.) We then apply the median-based recovery algorithm on only the remaining samples in the cells' interior. The interior *refined* sample points, by virtue of being far enough from their nearest cell boundary, would have remained in their respective cells, even after suffering from the rounding noise. Hence, we only have to ensure that all the interior regions have enough *good* samples, which we show increases the sample complexity by a factor dependent only on n. It still gives a tight upper bound on the sample complexity, in terms of d^n.

In order to avoid a run-time dependence on $(\|p\|_{\mathcal{C}_n,\infty}, \sigma)$, e.g., in case σ is unknown or $\|p\|_{\mathcal{C}_n,\infty}$ is too big, we compute an ℓ_1 minimizer $\widehat{p}_{\ell_1}$ first as in KKP's approach, However, for this analysis, we need a multivariate analog of Lemma 1.11 for the ℓ_1 norm. The main difficulty here is the fact that now we have many more 'peripheral' cells, i.e, cells on the boundary of $\mathcal{C}_n$ (these cells correspond to the peripheral intervals I_1 and I_m from the 1-dimensional Chebyshev partition, that needed Markov Brothers' Inequality). Since these peripheral cells are narrower, and much more in number, as n grows, this issue becomes more crucial. For example, for $n = 1$, the fraction of 'peripheral' intervals is $2/m$; but for $n = 2$, it is $\frac{4(m-1)}{m^2} = \frac{2}{m}(2 - 2/m) \gg \frac{1}{m^2}$. We circumvent this difficulty with our *second new technical contribution* (Theorem 1.9), that relates the ℓ_∞ and ℓ_1 norms of any individual degree-d, n-variate polynomial.

Relating ℓ_∞ and ℓ_1 norms of p. We inductively show the existence of a subset of points in $\mathcal{C}_n$, with a large measure (at least $1/(2d^2)^n$), on which the valuations of p can be guaranteed to be large, i.e., at least $\max_{\mathbf{x}\in\mathcal{C}_n} |p(\mathbf{x})|/2^n$. Thus we lower bound the ℓ_1 norm of p by a $1/\mathsf{poly}(d^n)$ factor of its ℓ_∞ norm, in the form of Theorem 1.9. We also note the tightness of this bound, by showing a family of polynomials for which their (resp.) ℓ_1 norms are upper bounded by a matching $1/\mathsf{poly}(d^n)$ factor of the (resp.) ℓ_∞ norms, in the form of Proposition 1.10.

To begin, for any point in $\mathbf{x} \in \mathcal{C}_n$, using Markov Brothers' inequality, we show the existence of a long enough line segment, on an axis-parallel line passing through it, such that p on that line segment has all valuations at least $p(\mathbf{x})/2$. For constructing (higher) $(k+1)$-dimensional cubes from k-dimensional cubes, in the induction step, we prove that all new unique line segments (one each corresponding to every point in the k-dimensional cube) can be translated to form a $(k+1)$-dimensional cube with a large enough Lebesgue measure. Thus, a sizable subset of points in n dimensions is constructed. On each of these points, the valuations of p are at least half of the valuations of p on the corresponding points in the k-dimensional cube. Using the inductive hypothesis, we conclude the argument.

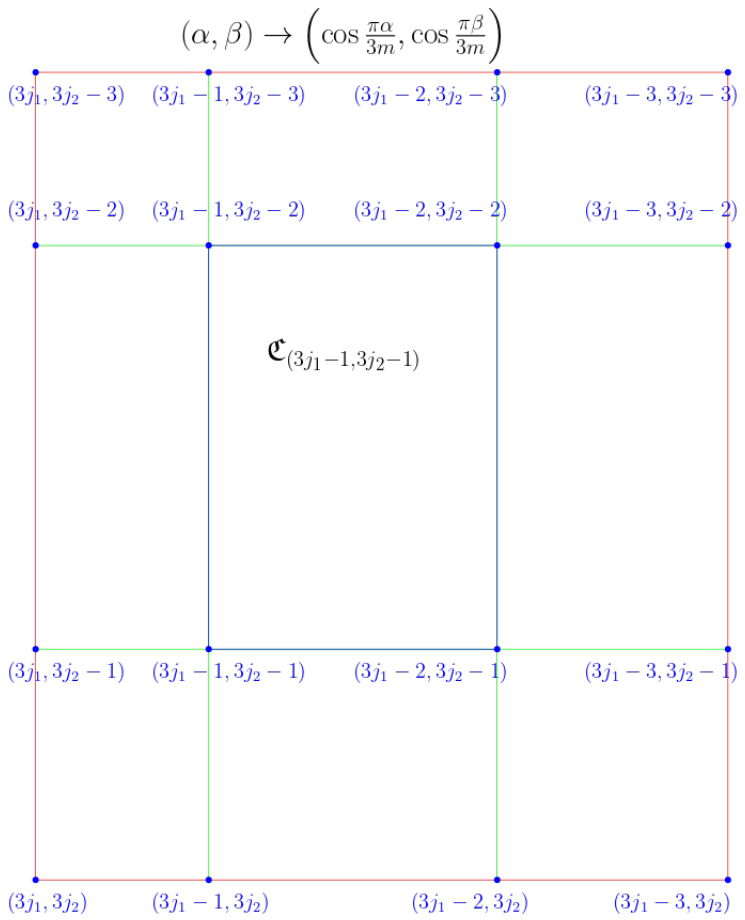

Figure 2: An illustration of cell-refinement in 2-dimensional Chebyshev grids: a $(3m, 2)$-grid (in green) super-imposed on a $(m, 2)$-Chebyshev cell $\mathcal{C}_{(j_1,j_2)}$ (in red). The samples from middle-most cell $\mathfrak{C}_{(3j_1-1,3j_2-1)}$ (in blue) only are retained, and median-recovery is applied on them.

Using Theorem 1.9 we bound the ℓ_∞ error of the ℓ_1 minimizer $\widehat{p}_{\ell_1}$ by $\mathsf{poly}(d^n)\sigma$. We then feed $\widehat{p}_{\ell_1}$ to the median based recovery procedure, which in $O(n \log d)$ iterations[5], brings down the error to $O(\sigma)$, thus proving Theorem 1.3.

Organization. We begin by setting up some preliminaries in Section 2. Discussion of the upper bounds follows, with Theorem 1.1 in Section 3, followed by Theorem 1.2 in Section 4. They rely on Theorem 1.8, which is proved in Section 3.1. Section 5 is devoted to proving ℓ_∞ and ℓ_1 norms relations, in the form of Theorem 1.9 and Proposition 1.10. Among them Theorem 1.9 is used to prove Theorem 1.3 in Section 6. Lower bounds are discussed in Section 7.

Acknowledgements

The authors would like to thank Yuval Filmus for fruitful discussions about some aspects of the robust regression problem.

[5]The number of iterations in this case additionally depends on $O(\log(1-2\rho))$, which diverges as $\rho \to 0.5$, thus agreeing with $\rho < 0.5$ being information-theoretically necessary.

2 Preliminaries

Notations. As mentioned earlier, $\mathcal{P}_d$ denotes the class individual degree-d polynomials, where a polynomial $p\colon \mathbb{R}^n \to \mathbb{R}$ is said to be of individual degree-d if it can be written as

$$p(x_1,\ldots,x_n) = \sum_{\boldsymbol{\alpha}\in\{0,1,\ldots,d\}^n} c_{\boldsymbol{\alpha}} x_1^{\boldsymbol{\alpha}_1}\cdots x_n^{\boldsymbol{\alpha}_n},$$

for some set of coefficients $c_{\boldsymbol{\alpha}} \in \mathbb{R}$. We use $[m] = \{1,\ldots,m\}$, and bold font for multi-index. For example $\boldsymbol{j} = (j_1,\ldots,j_n) \in [m]^n$ where each entry $j_i \in [m]$. We use the math bold font $\mathbf{x}, \mathbf{y}$ for vectors. To denote random uniform sampling from a set $\mathcal{D}$, we use $\sim \mathcal{D}$. We denote by $\mathcal{C}_n = [-1,1]^n$ the n dimensional solid cube and omit the subscript n when it is clear from the context. Our main problem of interest is the Robust Multivariate Polynomial Regression Problem, formally described in Section 1.

Definition 2.1. *[Norms] For any bounded subset $S \subsetneq \mathbb{R}^n$, for any $1 \le q \in \mathbb{R}$, the ℓ_q norm of a function $f : \mathbb{R}^n \to \mathbb{R}$ on S, provided it exists, is defined as:*

$$\|f\|_{S,q} \triangleq \left(\int_S |f(\mathbf{x})|^q d\mathbf{x}\right)^{\frac{1}{q}} < \infty.$$

The supremum norm of f on S is defined as $\|f\|_{S,\infty} \triangleq \lim_{q\to\infty}\|f\|_{S,q} = \sup_{\mathbf{x}\in S}\{|f(\mathbf{x})|\}$.

Lemma 2.2 (Markov Brothers' Inequality [Mar90])**.** *Let $p : \mathbb{R} \to \mathbb{R}$ be a degree-d polynomial. Then, for all $a < b \in \mathbb{R}$,*

$$\|p'\|_{[a,b],\infty} \le \frac{2d^2}{b-a}\|p\|_{[a,b],\infty}.$$

Chebyshev Partition. We partition the cube $\mathcal{C}_n$ into m^n cells by tensorizing the partition used by KKP for the line segment $[-1,1]$.

Definition 2.3 (Chebyshev partition)**.** *The (m,n)-Chebyshev partition of the cube $\mathcal{C}$ is a set of m^n cells indexed by $\boldsymbol{j} \in [m]^n$ and denoted $\mathcal{C}_{\boldsymbol{j}}$, such that*

$$\mathcal{C}_{\boldsymbol{j}} = \left[\cos\frac{\pi \boldsymbol{j}_1}{m}, \cos\frac{\pi(\boldsymbol{j}_1-1)}{m}\right] \times \cdots \times \left[\cos\frac{\pi \boldsymbol{j}_n}{m}, \cos\frac{\pi(\boldsymbol{j}_n-1)}{m}\right].$$

The grid is induced by partitioning $[-1,1]$ between the extrema points of the degree m Chebyshev polynomial[6] of the first kind, T_m, simultaneously along each axis.

We generalize KKP's notion of *goodness* that restricts the number of outliers in each cell:

Definition 2.4 (α-good sample set)**.** *We say that a set of samples $S = \{(\mathbf{x}_i, y_i)\}$ is α-good for the (m,n) Chebyshev partition, if for every $\boldsymbol{j} \in [m]^n$, the fraction of outliers in the cell $\mathcal{C}_{\boldsymbol{j}}$ is less than α.*

3 Main algorithmic result

In this section, we present the algorithm that solves the Robust Multivariate Polynomial Regression Problem, proving the following theorem, handling an approximation factor as close to 2 as we want, and any success probability $1-\delta$.

[6] see Definition 7.1, and Definition 7.2

Theorem 3.1. *[Generalized version of Theorem 1.1] Let $\varepsilon \in (0, 1/2], \delta \in (0, \varepsilon], \sigma \geq 0, \eta > 0$, and $\rho < 1/2$. There is an algorithm that almost solves the Robust Multivariate Polynomial Regression Problem up to an additive error of η. The output of the algorithm is a polynomial $\widehat{p}$ of degree at most d in each variable, such that with probability at least $1 - \delta$ (over the random input samples), $\widehat{p}$ satisfies*

$$|p(\mathbf{x}) - \widehat{p}(\mathbf{x})| \leq (2 + \varepsilon)\sigma + \eta \qquad \text{for all } \mathbf{x} \in \mathcal{C}\,.$$

It uses $M = O_{n,\rho}((d/\varepsilon)^n \log(d/\delta))$ samples drawn from the multidimensional Chebyshev distribution, or $M = O_{n,\rho}((d/\varepsilon)^{2n} \log(d/\delta))$ if the samples are drawn from the uniform measure. Its run-time is that of solving $O(\log_{1/\varepsilon}(\|p\|_{\mathcal{C}_n,\infty}/\eta)$ linear programs with $(d+1)^n < M$ variables, and M constraints.

Remark 3.2. *One may consider the case of non-constant values of $\rho < 1/2$. Here, the number of samples increases as $\rho \to 1/2$, since the dependence of M on ρ is $M \propto 1/(1-2\rho)^2$.*

We remark that the idea is to show that we may achieve a multiplicative approximation factor C, as close to 2 as we want (as long as $C > 2$), at the cost of more samples. We may allow larger values of ε, and then run our algorithm[7] with $\varepsilon' = \min\{\varepsilon, 1/2\}$. For $\varepsilon \geq 1/2$, the dependence on ε in the sample complexity becomes constant for constant values of n.

Remark 3.3. *In case $\sigma > 0$ is known, one may choose $\eta = \varepsilon\sigma/2$, and set the ε parameter to be half of the desired bound to guarantee $\|\widehat{p} - p\|_{\mathcal{C}_n,\infty} \leq (2 + \varepsilon)\sigma$.*

Our algorithm, given in Algorithm 2 with its subroutine Algorithm 1, is essentially the same algorithm proposed by KKP, which now uses the (m, n)-Chebyshev partition of the cube $\mathcal{C}$ instead of the $(m, 1)$-Chebyshev partition of the interval $[-1, 1]$ used in KKP. Compared to their algorithm, we don't use the ℓ_1 regression as the first step, but instead start with the 0 polynomial as the first approximation. We first describe the idea of the algorithm and the median-based recovery.

Median-based Recovery: As in KKP (a similar approach was taken by [DDL18]), for every $\boldsymbol{j} \in [m]^n$, we take the median $\tilde{y}_{\boldsymbol{j}}$ of all the y_i's corresponding to locations $\mathbf{x}_i$'s that land in the cell $\mathcal{C}_{\boldsymbol{j}}$. We assume that the sample set S is α-good, so the fraction of outliers in each cell is strictly less than one-half (since $\alpha < 1/2$) so that the median lies in between the values of the inlier labels. However, $\tilde{y}_{\boldsymbol{j}}$ may not itself be an inlier label for some sampled location $\mathbf{x}_i$. We generalize KKP's techniques to show that fitting an arbitrary $\tilde{\mathbf{x}}_{\boldsymbol{j}} \in \mathcal{C}_{\boldsymbol{j}}$ to the label $\tilde{y}_{\boldsymbol{j}}$ yields a good algorithm. We compute the polynomial r, that minimizes $\max_{\boldsymbol{j} \in [m]^n} |r(\tilde{\mathbf{x}}_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}|$, and show that r is $O(\sigma)$-close to p in ℓ_∞ up to an additive error of $\varepsilon\|p\|_{\mathcal{C}_n,\infty}$. To deal with this error, we iteratively refine the estimate r. After $\log_{1/\varepsilon}(\|p\|_{\mathcal{C}_n,\infty}/\eta)$ iterations, the additive error becomes as small as η.

To prove Theorem 3.1, we show that for M as in the theorem, the set of samples is α-good with high probability, and then we apply the following result.

Theorem 3.4 (Absolute ℓ_∞ error bound)**.** *Let c be some absolute constant, and let $\varepsilon, \alpha < 1/2$, $0 < \eta \leq 1$, be parameters. For any $m \geq cdn/\varepsilon$, if the set $S = \{(\mathbf{x}_i, y_i)\}$ of M samples is α-good for the (m, n)-Chebyshev partition, then the median-based recovery Algorithm 2 returns an individual degree-d polynomial $\widehat{p} = \widehat{p}^{(N_2)}$, such that*

$$\|p - \widehat{p}\|_{\mathcal{C}_n,\infty} \leq (2 + \epsilon)\sigma + \eta.$$

The first part of the proof of Theorem 3.4 follows the skeleton of the proof of [KKP17, Theorem 1.4], whilst skipping the ℓ_1 intermediate regression.

The main ingredient for proving Theorem 3.4 is the following technical result, bounding the ℓ_∞ error of the non-robust ℓ_∞ minimizer, i.e., a single run of Algorithm 1. This is later used to bound the error of the robust minimizer Algorithm 2.

[7] Having $\varepsilon' \leq 1/2$ is a limitation of the current analysis. An open question remains to make it work efficiently for any $\varepsilon' > 0$.

Algorithm 1: Refinement

Procedure REFINE($S, \widehat{p}$)
  Given : A set of samples $S = \{\mathbf{x}_i, y_i\}_{i=1}^M$, and an estimate $\widehat{p}$.
  for $\boldsymbol{j} \in [m]^n$ **do**
    $\tilde{y}_{\boldsymbol{j}} \leftarrow \mathsf{med}_{\mathbf{x}_i \in \mathcal{C}_{\boldsymbol{j}}}(y_i - \widehat{p}(\mathbf{x}_i))$;
    Choose an arbitrary $\tilde{\mathbf{x}}_{\boldsymbol{j}} \in \mathcal{C}_{\boldsymbol{j}}$;
  Fit a degree d polynomial r minimizing $\|r(\tilde{\mathbf{x}}_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}\|_\infty$;
  $\widehat{p}' \leftarrow \widehat{p} + r$;
  Return $\widehat{p}'$.

Algorithm 2: Median Based Recovery

Given : A set of samples $S = \{\mathbf{x}_i, y_i\}_{i=1}^M$, approximation factor $\varepsilon \le 1/2$. accuracy parameter $\eta > 0$.
$\widehat{p}^{(1)} \leftarrow$ REFINE$(S, 0)$; ▷ Let $\widehat{p}^{(1)}(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \{0,1,\ldots,d\}^n} c_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$
Let $v_{\max}$ be such that $|\widehat{p}^{(1)}(\mathbf{x})| \le v_{\max}$ for all $\mathbf{x} \in \mathcal{C}$; ▷ Set $v_{\max} \triangleq \sum_{\boldsymbol{\alpha} \in \{0,1,\ldots,d\}^n} |c_{\boldsymbol{\alpha}}|$
$N_2 \leftarrow O\left(\log_{1/\varepsilon}(v_{\max}/\eta)\right)$;
for $i \in \{1, \ldots, N_2 - 1\}$ **do**
  $\widehat{p}^{(i+1)} \leftarrow$ REFINE$(S, \widehat{p}^{(i)})$;
Return $\widehat{p}^{(N_2)}$.

Lemma 3.5 (Relative ℓ_∞ error bound, generalization of [KKP17, Lemma 1.3])**.** *Let $c > 0$ be an absolute constant. Let $\varepsilon, \alpha < 1/2$, and $m \ge cdn/\varepsilon$. Let the set $S = \{(\mathbf{x}_i, y_i)\}$ of M samples is α-good for the (m, n)-Chebyshev partition. And for every $\boldsymbol{j} \in [m]^n$, let $\tilde{\mathbf{x}}_{\boldsymbol{j}}$ be an arbitrary point from the cell $\mathcal{C}_{\boldsymbol{j}}$, and $\tilde{y}_{\boldsymbol{j}} \triangleq \mathsf{med}_S\{y_i : \mathbf{x}_i \in \mathcal{C}_{\boldsymbol{j}}\}$, i.e. the median of all those y_i's in S, whose corresponding $\mathbf{x}_i$ is in the cell $\mathcal{C}_{\boldsymbol{j}}$. Then, with*

$$\widehat{p} \triangleq \arg\min_{q \in \mathcal{P}_d} \max_{\boldsymbol{j} \in [m]^n} |q(\tilde{\mathbf{x}}_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}|, \tag{1}$$

where the minimization is over the class $\mathcal{P}_d$ of all individual degree-d polynomials over $\mathbb{R}^n$, we have

$$\|p - \widehat{p}\|_{\mathcal{C}_n,\infty} \le (2 + \varepsilon)\sigma + \varepsilon \|p\|_{\mathcal{C}_n,\infty}.$$

The proof of this statement mirrors the proof of its univariate counterpart [KKP17, Lemma 1.3], with Theorem 1.8 (which we prove in Section 3.1) replacing Lemma 1.11. Hence it is deferred to Appendix B.

Proof of Theorem 3.4. Let $\widehat{p}^{(t)}$ be the individual degree-d polynomial which is the t^{th} estimate of p computed by Algorithm 2, $e_t \triangleq p - \widehat{p}^{(t)}$ be the t^{th} error polynomial, where we define $\widehat{p}^{(0)} \equiv 0$, and $r_t(\mathbf{x}) \triangleq \arg\min_{r \in \mathcal{P}_d} \|r(\tilde{\mathbf{x}}_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}\|_\infty$, for $\tilde{y}_{\boldsymbol{j}} \triangleq \mathsf{med}_S\{y_i - \widehat{p}^{(t)}(\mathbf{x}_i) : \mathbf{x}_i \in \mathcal{C}_{\boldsymbol{j}}\}$, such that $\widehat{p}^{(t+1)} = \widehat{p}^{(t)} + r_t$. Note that for every inlier sample $i \in [M]$, we have $|e_t(\mathbf{x}_i) - (y_i - \widehat{p}^{(t)}(\mathbf{x}_i))| = |p(\mathbf{x}_i) - y_i| \le \sigma$, and therefore $S_t \triangleq \{(\mathbf{x}_i, y_i - \widehat{p}^{(t)}(\mathbf{x}_i))\}_{i=1}^M$ is α-good for e_t. So, by Lemma 3.5,

$$\|r_t - e_t\|_{\mathcal{C}_n,\infty} \le (2 + \varepsilon)\sigma + \varepsilon \|e_t\|_{\mathcal{C}_n,\infty}.$$

For every t, we have $r_t - e_t = (\widehat{p}^{(t+1)} - \widehat{p}^{(t)}) - (p - \widehat{p}^{(t)}) = \widehat{p}^{(t+1)} - p = -e_{t+1}$. So, we may deduce the inductive relation:

$$\|e_{t+1}\|_{\mathcal{C}_n,\infty} \le (2 + \varepsilon)\sigma + \varepsilon \|e_t\|_{\mathcal{C}_n,\infty}.$$

We use this relation to bound the error at step t:

$$\begin{aligned}\|p-\widehat{p}^{(t)}\|_{\mathcal{C}_n,\infty} &= \|e_t\|_{\mathcal{C}_n,\infty}\\ &\le (2+\varepsilon)(1+\varepsilon+\ldots+\varepsilon^t)\sigma+\varepsilon^t\|e_0\|_{\mathcal{C}_n,\infty}\\ &\le (2+6\varepsilon)\sigma+\varepsilon^t\|p\|_{\mathcal{C}_n,\infty}. \end{aligned} \tag{2}$$

The last inequality is from $\sum_{i\ge 0}\varepsilon^i \le 1/(1-\varepsilon) \le 1+2\varepsilon$ for $\varepsilon \le 1/2$, as well as $e_0 \equiv p$. So, for any $t \ge \log_{1/\varepsilon}(\|p\|_{\mathcal{C}_n,\infty}/\eta)$ we have

$$\|p-\widehat{p}^{(t)}\|_{\mathcal{C}_n,\infty} \le (2+6\varepsilon)\sigma+\eta.$$

Note that, after one iteration we already have $\widehat{p}^{(1)}$, an approximation of p, which might not be the best approximation, but we can still learn some bound on $\|p\|_{\mathcal{C}_n,\infty}$ from it. We will show that $N_2 = O\left(\log_{1/\varepsilon}(v_{\max}/\eta)\right)$ iterations, for $v_{\max} = \sum_{\boldsymbol{\alpha}\in\{0,1,\ldots,d\}^n}|c_{\boldsymbol{\alpha}}| \ge \|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}$, where $c_{\boldsymbol{\alpha}}$'s are the coefficients[8] of the polynomial $\widehat{p}^{(1)}$, suffice. By Lemma 3.5 and the triangle inequality, after the first iteration, we have

$$\|\widehat{p}^{(1)}-p\|_{\mathcal{C}_n,\infty} \le (2+\varepsilon)\sigma+\varepsilon\|p\|_{\mathcal{C}_n,\infty} \le (2+\varepsilon)\sigma+\varepsilon(\|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}+\|\widehat{p}^{(1)}-p\|_{\mathcal{C}_n,\infty}).$$

Rearranging and using $\varepsilon \le 1/2$, we get

$$\|\widehat{p}^{(1)}-p\|_{\mathcal{C}_n,\infty} \le \frac{1}{1-\varepsilon}((2+\varepsilon)\sigma+\varepsilon\|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}) \le 5\sigma+\|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}.$$

Again using the triangle inequality, we conclude:

$$\|p\|_{\mathcal{C}_n,\infty} \le \|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}+\|\widehat{p}^{(1)}-p\|_{\mathcal{C}_n,\infty} \le 5\sigma+2\|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}.$$

Plugging this into (2), we get

$$\begin{aligned}\|p-\widehat{p}^{(t)}\|_{\mathcal{C}_n,\infty} &\le (2+6\varepsilon)\sigma+\varepsilon^t(5\sigma+2\|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty})\\ &\le (2+6\varepsilon)\sigma+\varepsilon^t(5\sigma+2v_{\max}).\end{aligned}$$

Set $N_2 = \log_{1/\varepsilon}((5+2v_{\max})/\eta)+1$. Then if $\sigma \le 1$, we have $N_2 \ge \log_{1/\varepsilon}(\|p\|_{\mathcal{C}_n,\infty}/\eta)$, and for $t = N_2$ we have the desired approximation, after a further rescaling of ε to $\varepsilon/6$. Otherwise, for $\sigma \ge 1$ we have

$$\|p\|_{\mathcal{C}_n,\infty}/\sigma \le 5+2\|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}/\sigma \le 5+2\|\widehat{p}^{(1)}\|_{\mathcal{C}_n,\infty}.$$

Setting $N_2 \ge \log_{1/\varepsilon}(\|p\|_{\mathcal{C}_n,\infty}/\sigma)+1$ results in additive error at most $\varepsilon\sigma$. Thus, for $t = N_2$ we have,

$$\|p-\widehat{p}^{(t)}\|_{\mathcal{C}_n,\infty} \le (2+6\varepsilon)\sigma+\varepsilon\sigma = (2+7\varepsilon)\sigma.$$

Rescaling of ε to $\varepsilon/7$ gives the desired bound on the regression error with the additive error $= \eta > 0$. □

Before proving Theorem 3.1, we note a useful result from the analysis of multivariate functions.

Theorem 3.6 (Fubini-Tonelli Theorem (Theorem 14.2, [DiB02]))**.** *For any $X, Y \subseteq \mathbb{R}$, and some measurable $f : X \times Y \to \mathbb{R}_{\ge 0}$,*

$$\int_{X\times Y} f(x,y)d(x,y) = \int_Y\left(\int_X f(x,y)dx\right)dy = \int_X\left(\int_Y f(x,y)dy\right)dx.$$

Furthermore, if $f(x,y) \equiv f_1(x)f_2(y)$, for some measurable $f_1 : X \to \mathbb{R}_{\ge 0}$, and $f_2 : Y \to \mathbb{R}_{\ge 0}$, then

$$\int_{X\times Y} f(x,y)d(x,y) = \left(\int_X f_1(x)dx\right)\left(\int_Y f_2(y)dy\right).$$

[8] Given the coefficient representation of $\widehat{p}^{(1)}$ from the first iteration, $v_{\max}$ can be computed efficiently.

We also note a lower bound on the width of Chebyshev grid cells.

Observation 3.7. *For a cell $\mathcal{C}_{\boldsymbol{j}}$, denote its side length on direction t by $\mathcal{C}_{\boldsymbol{j}}(t)$ then*

$$|\mathcal{C}_{\boldsymbol{j}}(t)| = \left|\cos\frac{\pi \boldsymbol{j}_t}{m} - \cos\frac{\pi(\boldsymbol{j}_t - 1)}{m}\right| = \left|2\sin\frac{\pi}{2m}\sin\frac{\pi(2\boldsymbol{j}_t - 1)}{2m}\right| \geq \left(\frac{\pi}{2m}\right)^2. \tag{3}$$

The first equality follows by the trigonometric identity $\cos\theta - \cos\varphi = -2\sin((\theta+\varphi)/2)\sin((\theta-\varphi)/2)$. *The next inequality holds since* $\sin\theta \geq \theta/2$ *for all* $0 \leq \theta \leq \pi/2$ *(follows from the Taylor approximation* $\sin\theta \geq \theta - \theta^3/6$ *for* $\theta \geq 0$*). For the first factor, it is used with* $\theta = \frac{\pi}{2m}$ *and for the second factor with* $\theta = \frac{\pi}{2m}\min\{2\boldsymbol{j}_t - 1, 2(m+1-\boldsymbol{j}_t) - 1\}$ *noting that* $\sin\frac{\pi(2\boldsymbol{j}_t-1)}{2m} = \sin\frac{\pi(2(m+1-\boldsymbol{j}_t)-1)}{2m}$. *We also use* $1 \leq \min\{2\boldsymbol{j}_t - 1, 2(m+1-\boldsymbol{j}_t) - 1\} \leq m$.

Proof of Theorem 3.1. We need to show that the input set of samples S is α-good for the (m,n)-Chebyshev partition for some $\alpha < 1/2$ with probability at least $1-\delta$. Then for $m = cdn/\varepsilon$, for some constant $c > 0$, the result follows from Theorem 3.4. Consider the (m,n)-Chebyshev partition. For any $\boldsymbol{j} \in [m]^n$, let $p_{\boldsymbol{j}}$ be the probability that a sample from χ is in $\mathcal{C}_{\boldsymbol{j}}$. Let $X_{\boldsymbol{j}}$ be the number of samples in $\mathcal{C}_{\boldsymbol{j}}$, and $Y_{\boldsymbol{j}}$ be the number of outliers in $\mathcal{C}_{\boldsymbol{j}}$.

The probability that S violates the α-goodness for a cell $\mathcal{C}_{\boldsymbol{j}}$ can be bounded by the law of total probability,

$$\begin{aligned}\Pr[Y_{\boldsymbol{j}} \geq \alpha X_{\boldsymbol{j}}] \leq \Pr[Y_{\boldsymbol{j}} \geq \alpha X_{\boldsymbol{j}} | X_{\boldsymbol{j}} \leq Mp_{\boldsymbol{j}}/2] \underbrace{\Pr[X_{\boldsymbol{j}} \leq Mp_{\boldsymbol{j}}/2]}_{(I)} \\ + \underbrace{\Pr[Y_{\boldsymbol{j}} \geq \alpha X_{\boldsymbol{j}} | X_{\boldsymbol{j}} > Mp_{\boldsymbol{j}}/2]}_{(II)} \Pr[X_{\boldsymbol{j}} > Mp_{\boldsymbol{j}}/2].\end{aligned}$$

Next, we bound each term separately. For the first term, note that $\mathbb{E}[X_{\boldsymbol{j}}] = Mp_{\boldsymbol{j}}$, so using Chernoff bound we have,

$$(I) \leq \Pr[|X_{\boldsymbol{j}} - \mathbb{E}[X_{\boldsymbol{j}}]| \geq Mp_{\boldsymbol{j}}/2] \leq 2e^{-Mp_{\boldsymbol{j}}/12}.$$

For the second term, $\mathbb{E}[Y_{\boldsymbol{j}} | X_{\boldsymbol{j}} = x] = \rho x$, and using Hoeffding's Inequality, we have

$$\Pr[Y_{\boldsymbol{j}} \geq \alpha X_{\boldsymbol{j}} \mid X_{\boldsymbol{j}}] \leq \Pr[|Y_{\boldsymbol{j}} - \mathbb{E}[Y_{\boldsymbol{j}}]| \geq (\alpha - \rho)X_{\boldsymbol{j}} \mid X_{\boldsymbol{j}}] \leq 2e^{-(\alpha-\rho)^2 X_{\boldsymbol{j}}}.$$

Setting $\alpha = \frac{2\rho+1}{4}$ we have $\alpha < \frac{1}{2}$, and $(\alpha-\rho)^2 = \left(\frac{1-2\rho}{4}\right)^2 = (1-2\rho)^2/16$, giving us

$$(II) \leq 2e^{-(1-2\rho)^2 X_{\boldsymbol{j}}/16} \leq 2e^{-(1-2\rho)^2 Mp_{\boldsymbol{j}}/32}.$$

The last inequality is by the condition $X_{\boldsymbol{j}} > Mp_{\boldsymbol{j}}/2$. We conclude that the failure probability is

$$\Pr[\exists \boldsymbol{j} \in [m]^n : Y_{\boldsymbol{j}} \geq \alpha X_{\boldsymbol{j}}] \leq \sum_{\boldsymbol{j}\in[m]^n} \Pr[Y_{\boldsymbol{j}} \geq \alpha X_{\boldsymbol{j}}] \leq \sum_{\boldsymbol{j}\in[m]^n} \left(2e^{-(1-2\rho)^2 Mp_{\boldsymbol{j}}/32} + 2e^{-Mp_{\boldsymbol{j}}/12}\right). \tag{4}$$

The first inequality is by a union bound over all the m^n cells. For the second inequality, we plugged in the bounds for $(I), (II)$. By Observation 3.7, we deduce that for any $\boldsymbol{j} \in [m]^n$, a point uniformly sampled from $\mathcal{C}$ falls into $\mathcal{C}_{\boldsymbol{j}}$ with probability

$$p_U(\mathcal{C}_{\boldsymbol{j}}) = \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{2^n} \geq \left(\frac{\pi^2}{8m^2}\right)^n \geq m^{-2n}.$$

For sampling from the unidimensional Chebyshev distribution, using the fact that $\int \frac{dx}{\sqrt{1-x^2}} = \arcsin(x)$, KKP observed: for any $j \in [m]$,

$$\int_{\cos(\pi j)/m}^{\cos(\pi(j-1)/m)} \frac{1}{\sqrt{1-x^2}} dx = \arcsin\left(\cos\frac{\pi(j-1)}{m}\right) - \arcsin\left(\cos\frac{\pi j}{m}\right) = \frac{\pi}{m}. \tag{5}$$

So, for sampling from the n-dimensional Chebyshev distribution, the probability that $\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}$ becomes

$$\begin{aligned}
p_C(\mathcal{C}_{\boldsymbol{j}}) &= \int_{\mathcal{C}_{\boldsymbol{j}}(n)} \dots \int_{\mathcal{C}_{\boldsymbol{j}}(1)} \frac{1}{\pi\sqrt{1-\mathbf{x}_1^2}} \times \dots \times \frac{1}{\pi\sqrt{1-\mathbf{x}_n^2}}\, d\mathbf{x}_1 \dots d\mathbf{x}_n && \text{(Integrands are all non-negative)} \\
&= \prod_{i=1}^{n} \left(\int_{\mathcal{C}_{\boldsymbol{j}}(i)} \frac{1}{\pi\sqrt{1-\mathbf{x}_i^2}} d\mathbf{x}_i \right) && \text{(Splitting independent integrals by Theorem 3.6)} \\
&= \prod_{i=1}^{n} \left(\frac{1}{\pi} \int_{\cos(\pi \boldsymbol{j}_i/m)}^{\cos(\pi(\boldsymbol{j}_i-1)/m)} \frac{1}{\sqrt{1-\mathbf{x}_i^2}} d\mathbf{x}_i \right) && (\because \mathcal{C}_{\boldsymbol{j}}(i) = [\cos(\pi \boldsymbol{j}_i/m), \cos(\pi(\boldsymbol{j}_i-1)/m)]) \\
&= \prod_{i=1}^{n} \frac{1}{m} = \frac{1}{m^n}. && \text{(By using (5) and since } \boldsymbol{j}_i \in [m], \forall i \in [n])
\end{aligned}$$

For Chebyshev sampling, with $p_{\boldsymbol{j}} = m^{-n}$, and upper bounding the failure probability from (4) by at most δ, we get S is α-good for

$$M_C(m) = (1-2\rho)^{-2} m^n \log(m^n/\delta). \tag{6}$$

For Algorithm 2, with $m = cdn/\varepsilon$, for some constant $c > 0$, this gives us a Chebyshev sample complexity of $M_C = \frac{1}{(1-2\rho)^2}(cnd/\varepsilon)^n \log\frac{d}{\varepsilon\delta}$. For uniform sampling, replacing $p_{\boldsymbol{j}}$ by the bound on $p_U(\mathcal{C}_{\boldsymbol{j}}) \geq m^{-2n}$ and then plugging in

$$M = M_U(m) = (1-2\rho)^{-2} m^{2n} \log(4m^n/\delta) \tag{7}$$

in (4), we get that S is α-good with probability:

$$\Pr[Y_{\boldsymbol{j}} < \alpha X_{\boldsymbol{j}}, \forall \boldsymbol{j}] \geq 1 - 4 \sum_{\boldsymbol{j} \in [m]} \frac{\delta}{4m^n} = 1 - \delta$$

Thus, with $m = cdn/\varepsilon$, the Uniform sample complexity is $M_U = \frac{1}{(1-2\rho)^2}(cnd/\varepsilon)^{2n} \log\frac{d}{\delta}$, for some constant $c > 0$. The run-time of Algorithm 2 is thus that of solving $N_2 = O(\log_{1/\varepsilon}(\|p\|_{\mathcal{C}_n,\infty}/\eta))$ linear programs with $O(d^n)$ variables, and M constraints. □

3.1 Approximating polynomials by piece-wise constant functions

In this section, we prove Theorem 1.8, by generalizing Lemma 1.11 for the case of ℓ_∞, to the multivariate setting. The idea here is to approximate p on a fine enough Chebyshev grid, by an arbitrary piece-wise constant function r, that is (i) constant on every Chebyshev cell, and (ii) consistent with p on some point in every cell. We show that the finer the grid is, the smaller the difference $p - r$ becomes, and hence, the better an approximation r may be for p. We first restate Lemma 1.11 as our generalization is through reducing to the univariate case.

Lemma 1.11. *[Lemma 2.1, [KKP17]] Let $g : \mathbb{R} \to \mathbb{R}$ be a (univariate) degree-d polynomial. Let $\{I_j\}_{j\in[m]}$ denote the m-Chebyshev partition of $[-1,1]$, for some $m \geq d$. Let $r : [-1,1] \to \mathbb{R}$ be piece-wise constant,*

so that for each $k \in [m]$, there exists $x_k^* \in I_k$, such that $r(x) = g(x_k^*)$ for all $x \in I_k$. Then, there exists a universal constant C such that, for any $q \geq 1$,

$$\|g - r\|_q \leq \frac{Cd}{m}\|g\|_q.$$

We prove how we can conclude from Lemma 1.11, a similar result for the multivariate case, where we replace the interval $[-1, 1]$ with the n dimensional solid cube $\mathcal{C} = [-1, 1]^n$. The polynomial $p\colon \mathcal{C} \to \mathbb{R}$ is now of individual degree at most d (in each variable), and the function r is now piece-wise constant with respect to the (m, n)-Chebyshev partition.

Theorem 1.8. *[Approximation by piece-wise constant functions] Let $p : \mathcal{C}_n \to \mathbb{R}$ be a polynomial of degree at most d in each variable, and $m \geq d$. Let $r : \mathcal{C}_n \to \mathbb{R}$ be a piece-wise constant function with respect to the (m, n)-Chebyshev partition, such that for every $\boldsymbol{j} \in [m]^n$, there exist a point $\mathbf{x}^{(\boldsymbol{j})} \in \mathcal{C}_{\boldsymbol{j}}$, such that $r(\mathbf{x}) = p(\mathbf{x}^{(\boldsymbol{j})})$, for all $\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}$. Then, there exists a universal constant C such that,*

$$\|p - r\|_{\mathcal{C}_n,\infty} \leq C\frac{dn}{m}\|p\|_{\mathcal{C}_n,\infty}.$$

Before we prove Theorem 1.8, we note and prove a useful technical claim. Though its proof is simple, it exposes the idea behind Lemma 1.11 (at least for the ℓ_∞ case).

Claim 3.8. *Let $g : [-1, 1] \to \mathbb{R}$ be a degree-d univariate polynomial. Consider the m-size Chebyshev partition of $[-1, 1]$, denoted by $I_1, \ldots, I_m$. For any $i \in [m]$, and any two points $\alpha, \beta \in I_i$, there exists a function $v\colon [-1, 1] \to \mathbb{R}$ that is piece-wise constant with respect to g, and the partition, such that:*

$$|g(\alpha) - g(\beta)| \leq \|g - v\|_{[-1,1],\infty}. \tag{8}$$

Furthermore, there exists a universal constant C such that $|g(\alpha) - g(\beta)| \leq \frac{Cd}{m}\|g\|_{[-1,1],\infty}$

Proof. Fix $i \in [m]$, and $\alpha, \beta \in I_i$. We define the piece-wise constant function v as follows,

- For any $j \in [m]$ such that $j \neq i$, choose an arbitrary point $x_j \in I_j$, and set $v(x) = g(x_j)$ for all $x \in I_j$.
- For I_i, set $v(x) = g(\beta)$ for all $x \in I_i$.

Note that, in particular, $v(\alpha) = g(\beta)$. Hence,

$$|g(\alpha) - g(\beta)| = |g(\alpha) - v(\alpha)| \leq \max_{\gamma\in[-1,1]} |g(\gamma) - v(\gamma)| = \|g - v\|_{[-1,1],\infty}.$$

By Lemma 1.11, we conclude $|g(\alpha) - g(\beta)| \leq \frac{Cd}{m}\|g\|_{[-1,1],\infty}$, for some universal constant C. □

We are now ready to prove Theorem 1.8:

Proof of Theorem 1.8. Observe,

$$\begin{aligned}
&\|p - r\|_{\mathcal{C}_n,\infty} && (9)\\
&\quad= \max_{\boldsymbol{j}\in[m]^n,\mathbf{x}\in\mathcal{C}_{\boldsymbol{j}}} |p(\mathbf{x}) - r(\mathbf{x})|\\
&\quad= \max_{\boldsymbol{j}\in[m]^n,\mathbf{x}\in\mathcal{C}_{\boldsymbol{j}}} |p(\mathbf{x}) - p(\mathbf{x}^{(\boldsymbol{j})})|\\
&\quad\leq \max_{\boldsymbol{j}\in[m]^n,\mathbf{x},\mathbf{y}\in\mathcal{C}_{\boldsymbol{j}}} |p(\mathbf{x}) - p(\mathbf{y})| && (10)
\end{aligned}$$

$$\leq \max_{\boldsymbol{j}\in[m]^n,\mathbf{x},\mathbf{y}\in\mathcal{C}_{\boldsymbol{j}}} \sum_{k=1}^{n} |p(x_1,\ldots,x_{k-1},x_k,y_{k+1},\ldots,y_n) - p(x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n)|,$$

where the last inequality is by a hybrid argument: walking from the point $\mathbf{x}$ to the point $\mathbf{y}$ along the axes, gives the successive summands in (9). We next bound each of the summands by $C\frac{d}{m}\|p\|_{\mathcal{C}_n,\infty}$, then summing up all n of them results with the desired bound on $\|p-r\|_{\mathcal{C}_n,\infty}$. For every $k\in[n]$, we observe that

$$(x_1,\ldots,x_{k-1},x_k,y_{k+1},\ldots,y_n), \text{ and } (x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n)$$

are points on the line $\mathrm{L}_{(x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_k}$.
So, they provide evaluations of $p_{\mathrm{L}_{(x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_k}}(t)$, the univariate line restriction polynomial, which has degree[9] at most d, since $p:\mathcal{C}\to\mathbb{R}$ has individual degree at most d, for every variable. And, since $\mathbf{x},\mathbf{y}\in\mathcal{C}_{\boldsymbol{j}}$, we have $x_k,y_k\in I_i$, for $i=\boldsymbol{j}_k\in[m]$. Thus, by Claim 3.8,

$$\begin{aligned}
&|p(x_1,\ldots,x_{k-1},x_k,y_{k+1},\ldots,y_n) - p(x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n)| \\
&\leq \frac{Cd}{m}\|p_{\mathrm{L}_{(x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_k}}\|_{[-1,1],\infty} \\
&\leq \frac{Cd}{m}\|p\|_{\mathcal{C}_n,\infty},
\end{aligned}$$

as needed. □

4 Dealing with finite precision representations

We prove a more precise statement of Theorem 1.2, giving an algorithm for handling an approximation factor close enough to 2, and for any success probability $1-\delta$.

Theorem 4.1. *[Generalized version of Theorem 1.2] Let N be the number of bits of precision, $\sigma\geq 2^{-N}$ and, constant $\rho<1/2$. For any $\varepsilon\leq 1/2$ such that $\varepsilon=\Omega_n(d2^{-N/2})$, and $\delta\in(0,\varepsilon]$, there exists an algorithm for the Robust Multivariate Polynomial Regression Problem. The output of the algorithm is $\widehat{p}\colon\mathbb{R}^n\to\mathbb{R}$, a polynomial of degree at most d in each variable, that satisfies*

$$|p(\mathbf{x})-\widehat{p}(\mathbf{x})|\leq(2+\varepsilon)\sigma \qquad \text{for all } \mathbf{x}\in\mathcal{C}_n,$$

with probability at least $1-\delta$. It uses $M=O_{n,\rho}((d/\varepsilon)^n\log(d/\delta))$ samples drawn from the multidimensional Chebyshev distribution, or $M=O_{n,\rho}((d/\varepsilon)^{2n}\log(d/\delta))$ if the samples are drawn from the uniform measure. Its run-time is that of solving $O(\log_{1/\varepsilon}\|p\|_{\mathcal{C}_n,\infty}+N)$ linear programs with $(d+1)^n<M$ variables, and M constraints.

Remark 4.2. *We note that the condition on ε implies that in order to learn degree d polynomials using Algorithm 2, one needs at least $N=\Omega(\log d)$ bits of precision. The reason is that to get a good approximation we need to take a fine enough grid, but the grid's "fineness parameter" m is limited as well by the precision restriction, as we need the width of any cell to be at least 2^{-N}. Note that if $N=o(\log d)$, i.e. $d=2^{\Omega(N)}$, then $\varepsilon=\Omega(1)$, i.e. the approximation factor achieved in this setting is too large, compared to the factor of at most 3, achievable when $N=\Omega(\log d)$.*

[9] The line $\mathrm{L}_{(x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_k}$ being *axis-parallel* is crucial here, allowing us to bound the degree of the line restriction polynomial $p_{\mathrm{L}_{(x_1,\ldots,x_{k-1},y_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_k}}(t)$ by the individual degree of p, as t varies only along $\boldsymbol{e}_k$.

Proof of Theorem 4.1. The algorithm that solves this problem, uses Algorithm 2 as a black box, with $\eta = \varepsilon 2^{-N}$. We note that if the samples $\mathbf{x}_i \in \mathcal{C}$ are exact, i.e. described using infinite precision, then we get the result by applying Theorem 3.1, and rescaling ε to $\varepsilon/2$.

Otherwise, we note that the only information used by Algorithm 2 from the sample set, for every $i \in [M]$ is: (i) the value y_i, and (ii) the index $\boldsymbol{k}_i \in [m]^n$ of the cell $\mathcal{C}_{\boldsymbol{k}_i}$ into which $\mathbf{x}_i$ lands. The inaccuracy[10] of (i) is covered by having $\sigma \geq 2^{-N}$. For (ii), we add a preliminary step to our algorithm, first sifting out the samples, that we don't know as to which cell $\mathcal{C}_{\boldsymbol{j}}$ they belong. These are the samples which are close to the borders of the cell, i.e. within a distance of 2^{-N} from any border of the cell. More precisely, for an input sample set $S = \{(\mathbf{x}_i, y_i)\}$ of size $|S| = M$, we run Algorithm 2 on the *restricted* sample set $S' \subseteq S$ defined as follows:

$$S' = \left\{(\mathbf{x}, y) \in S : \forall i \in [n].\ |k_{x_i} - x_i| > 2^{-N}\right\},$$

where k_{x_i} is the closest Chebyshev extrema to x_i. We note the samples we omit may come, originally, from another cell or from the same cell within an additional distance of 2^{-N}; the latter ones are good for us, but we have no way to distinguish between these two types. So, we omit them all.

We next show that even if we omit those samples, we only need to multiply the number of samples by a factor dependent only on n, to ensure we have enough samples for which we are sure as to which cell they belong. The sifting process is promised to save all the samples that, originally (i.e. in their exact infinite-bit precision representation), before the noise is applied, lie in the interior of the cell at a distance of at least $2 \cdot 2^{-N}$ (twice the precision noise) from any Chebyshev extrema (which determine the nearest cell boundary). This process requires the interior of the cell, when we omit the width $\sigma' = 2 \cdot 2^{-N}$ boundary, to exist. This restricts the partition to be coarse enough, i.e., we won't be able to take m to be too large, which means that ε cannot be too small. In particular, we need the side length of a cell $|\mathcal{C}_{\boldsymbol{j}}(i)| > 2\sigma'$.

Consider the interior of a cell $\mathcal{C}_{\boldsymbol{j}}$, denoted by $\mathcal{C}^*_{\boldsymbol{j}}$, defined as the region of $\mathcal{C}_{\boldsymbol{j}}$, within a boundary of width $\sigma' \triangleq \frac{1}{4}\left(\frac{\pi}{2m}\right)^2 = \left(\frac{\pi}{4m}\right)^2 \geq 2 \cdot 2^{-N}$, i.e. $|\mathcal{C}^*_{\boldsymbol{j}}(i)| = |\mathcal{C}_{\boldsymbol{j}}(i)| - 2\sigma'$, for every $i \in [n]$. By Observation 3.7, $|\mathcal{C}_{\boldsymbol{j}}(i)| \geq \left(\frac{\pi}{2m}\right)^2$, and hence, for every $i \in [n]$ we have,

$$|\mathcal{C}^*_{\boldsymbol{j}}(i)| \geq \left(\frac{\pi}{2m}\right)^2 - 2\sigma' \geq \frac{1}{2}\left(\frac{\pi}{2m}\right)^2 = \left(\frac{\pi}{8m}\right)^2.$$

So, for uniform sampling, we get

$$p_U(\mathcal{C}^*_{\boldsymbol{j}}) = \frac{V_n(\mathcal{C}^*_{\boldsymbol{j}})}{2^n} \geq \left(\frac{\pi^2}{16m^2}\right)^n \geq \left(\sqrt{2}m\right)^{-2n}.$$

Plugging this into (4), with $p_{\boldsymbol{j}}$ replaced by the bound on $p_U(\mathcal{C}^*_{\boldsymbol{j}})$, we get S is α-good for $M_U = M_U(\sqrt{2}m)$ samples, where $M_U(\cdot)$ is the formulation in (7). Then, with $m = c_1 nd/\varepsilon$ for some constant $c_1 > 0$, we get

$$\begin{aligned} M_U = M_U(\sqrt{2}m) &= (1-2\rho)^{-2} 2^n m^{2n} \log\left(4(\sqrt{2}m)^n/\delta\right) \\ &= \frac{1}{(1-2\rho)^2}\left(\frac{cnd}{\varepsilon}\right)^{2n} \log\frac{d}{\delta}, \end{aligned}$$

for some constant $c > 0$. Note, this is asymptotically the same as the M_U obtained for Theorem 3.1.

For Chebyshev sampling, we consider the $(3m, n)$-Chebyshev grid $\mathfrak{C} \triangleq \{\mathfrak{C}_{\boldsymbol{j}}, \boldsymbol{j} \in [3m]^n\}$, where

$$\mathfrak{C}_{\boldsymbol{j}} = \left[\cos\frac{\pi \boldsymbol{j}_1}{3m}, \cos\frac{\pi(\boldsymbol{j}_1 - 1)}{3m}\right] \times \cdots \times \left[\cos\frac{\pi \boldsymbol{j}_n}{3m}, \cos\frac{\pi(\boldsymbol{j}_n - 1)}{3m}\right].$$

We may observe, $\mathfrak{C}$ is a cell-refinement of the (m, n)-Chebyshev grid $\{\mathcal{C}_{\boldsymbol{j}}, \boldsymbol{j} \in [m]^n\}$. Formally:

[10] from finite-bit precision representations of reals

Claim 4.3. *For every $\boldsymbol{j} = (j_1, \ldots, j_n) \in [m]^n$, the cell $\mathcal{C}_{\boldsymbol{j}}$ contains all the 3^n refined cells $\mathfrak{C}_{(k_1,\ldots,k_n)}$, where $k_i \in \{3j_i, 3j_i - 1, 3j_i - 2\}$, for all $i \in [n]$, and nothing else.*

Proof. Consider some fixed $\boldsymbol{j} = (j_1, \ldots, j_n) \in [m]^n$, and the cell $\mathcal{C}_{\boldsymbol{j}}$. For every $i \in [n]$, its i^{th} side $\mathcal{C}_{\boldsymbol{j}}(i)$ is precisely the union of the i^{th} sides of all the 3^n refined cells contained in $\mathcal{C}_{\boldsymbol{j}}$:

$$\begin{aligned}
\mathcal{C}_{\boldsymbol{j}}(i) &= \left[\cos\left(\frac{\pi \boldsymbol{j}_i}{m}\right), \cos\left(\frac{\pi(\boldsymbol{j}_i - 1)}{m}\right)\right] \\
&= \left[\cos\left(\frac{\pi(3\boldsymbol{j}_i)}{3m}\right), \cos\left(\frac{\pi(3\boldsymbol{j}_i - 1)}{3m}\right)\right] \cup \left[\cos\left(\frac{\pi(3\boldsymbol{j}_i - 1)}{3m}\right), \cos\left(\frac{\pi(3\boldsymbol{j}_i - 2)}{3m}\right)\right] \\
&\quad \cup \left[\cos\left(\frac{\pi(3\boldsymbol{j}_i - 2)}{3m}\right), \cos\left(\frac{\pi(3\boldsymbol{j}_i - 3)}{3m}\right)\right] \quad \text{(By continuity, and monotonicity of } \cos \text{ on } \mathcal{C}_{\boldsymbol{j}}(i)) \\
&= \bigcup_{t=0}^{2} \left[\cos\left(\frac{\pi(3\boldsymbol{j}_i - t)}{3m}\right), \cos\left(\frac{\pi(3\boldsymbol{j}_i - (t+1))}{3m}\right)\right] = \bigcup_{t=0}^{2} \mathfrak{C}_{(\circledast,\ldots,\circledast,3\boldsymbol{j}_i - t,\circledast,\ldots,\circledast)}(i),
\end{aligned}$$

where each $\circledast_r$ (denoting $\circledast$ in the r^{th} position, for all $r \in [3m] \setminus \{i\}$), could be any of the $k \in [3m]$. The n relations, one for each side, then fix each $\circledast_r$ to be one of $\{3j_i, 3j_i - 1, 3j_i - 2\}$. □

We now want to ensure that, in every cell $\mathcal{C}_{\boldsymbol{j}}$, for every sample in the middle-most cell $\mathfrak{C}_{(3j_1-1,\ldots,3j_n-1)}$ the precision error doesn't move the sample to some other $\mathcal{C}_{\boldsymbol{j}'}$. So, we need

$$\sigma' \geq \max_{j_i \in [m]} \min\left\{\cos\frac{3j_i - 1}{3m} - \cos\frac{3j_i - 2}{3m}, \cos\frac{3j_i - 2}{3m} - \cos\frac{3j_i}{3m}\right\} \geq \left(\frac{\pi}{6m}\right)^2.$$

With this, to then ensure that $\mathfrak{C}_{(3j_1-1,\ldots,3j_n-1)}$ has sufficiently many good samples, we observe

$$p_C(\mathfrak{C}_{(3j_1-1,\ldots,3j_n-1)}) = \frac{1}{(3m)^n} = \frac{1}{3^n} p_C(\mathcal{C}_{\boldsymbol{j}}).$$

Again, plugging this into (4), with p_j replaced by $p_C(\mathcal{C}^*_{\boldsymbol{j}})$, we get S is α-good for $M_C = M_C(3m)$ samples, where $M_C(\cdot)$ is the formulation in (6). With $m = c_1 nd/\varepsilon$ for some constant $c_1 > 0$, we get

$$\begin{aligned}
M_C = M_C(3m) &= (1 - 2\rho)^{-2}(3m)^n \log((3m)^n/\delta) \\
&= \frac{1}{(1 - 2\rho)^2}\left(\frac{cnd}{\varepsilon}\right)^n \log\frac{d}{\delta},
\end{aligned}$$

for some constant $c > 0$. Note, this too is asymptotically the same as the M_C obtained for Theorem 3.1.

So, we can handle $\sigma = 2^{-N} \leq \left(\frac{\pi}{6m}\right)^2$. We need $m \geq c \max\{2^{N/2}, \frac{dn}{\varepsilon}\}$. Let $\sigma = 2^{-N}$ for some fixed N, then $\sigma' \geq \sigma$ implies $m \leq (\pi/6)2^{N/2}$. Simultaneously, we need $m > cdn/\varepsilon$. This induce the condition on $\varepsilon \geq \frac{cdn}{m} \geq c_0 dn 2^{-N/2}$ (for $c_0 = 6c/\pi$). □

5 From ℓ_∞ error to ℓ_1 error

In this independent section, we prove our second technical new idea, restated here for convenience:

Theorem 1.9. *There exists a global constant $C > 0$ such that, for every polynomial $p \in \mathcal{P}_d$:*

$$\|p\|_{\mathcal{C}_n,\infty} \leq C^n d^{2n} \|p\|_{\mathcal{C}_n,1}.$$

This, in Section 6, allows us to design and analyze a variant of Algorithm 2, that has a run-time independent of $\|p\|_{\mathcal{C}_n,\infty}$. To prove Theorem 1.9, we need the following result about univariate polynomials, which we next generalize for the multivariate case.

Lemma 5.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a degree at most d polynomial. Let $x^* \in [-1,1]$ be such that $|g(x^*)| = \|g\|_{[-1,1],\infty}$. Consider the set $S \triangleq \{x \in [-1,1] : |x^* - x| \leq 1/2d^2\}$. Then, for every $x \in S$,*

$$|g(x)| \geq |g(x^*)|/2.$$

To prove it, we need the following fundamental result from the analysis of univariate functions.

Theorem 5.2 (Mean Value Theorem (Theorem 5.10, [Rud76]))**.** *Let $a < b \in \mathbb{R}$, and $f : [a,b] \to \mathbb{R}$ be continuous on $[a,b]$, and differentiable on (a,b). Then there exists a point $z \in (a,b)$ such that*

$$f(b) - f(a) = (b-a)f'(z).$$

Proof of Lemma 5.1. Fix some $x \in S$. By Theorem 5.2, there exists a z between x^* and x, such that $g(x^*) - g(x) = g'(z)(x^* - x)$. Since $z \in [-1,1]$, by Lemma 2.2, we have $|g'(z)| \leq d^2\|g\|_{[-1,1],\infty}$. So,

$$|g(x^*) - g(x)| \leq d^2\|g\|_{[-1,1],\infty}|x^* - x| \leq \frac{\|g\|_{[-1,1],\infty}}{2} = \frac{|g(x^*)|}{2}.$$

The second inequality follows since $|x^* - x| \leq 1/2d^2$, for all $x \in S$. By the triangle inequality, we conclude

$$|g(x)| \geq |g(x^*)| - |g(x^*) - g(x)| \geq |g(x^*)|/2. \qquad \square$$

Using this, we can prove a local Lipschitz-like argument along *axis-parallel* lines, for any point in $\mathcal{C}$:

Corollary 5.3. *Let $p\colon \mathcal{C} \to \mathbb{R}$ be a polynomial of individual degree at most d. For every $\boldsymbol{a} \in \mathcal{C}$, and $i \in [n]$, consider the **axis-parallel** line $\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \triangleq \{\boldsymbol{a} + t\boldsymbol{e}_i : t \in \mathbb{R}\}$, passing through $\boldsymbol{a}$.*
There exists a line segment $J \subset \mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \cap \mathcal{C}$ of length at least $\frac{1}{2d^2}$, such that for every $\mathbf{x} \in J$, $|p(\mathbf{x})| \geq \frac{|p(\boldsymbol{a})|}{2}$.

Proof. Fix $\boldsymbol{a} \in \mathcal{C}$ and $i \in [n]$. Observe, p restricted to $\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}$, $p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}(t) \triangleq p(\boldsymbol{a} + t\boldsymbol{e}_i)$, i.e. we fix $x_j = \boldsymbol{a}_j$ for all $j \neq i$, and only let x_i vary. It is a univariate polynomial in the formal variable t, of degree[11] at most d. Consider a point $\mathbf{x}^* \in \mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \cap \mathcal{C}$ such that $|p(\mathbf{x}^*)| = \|p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}\|_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \cap \mathcal{C}_n,\infty}$. For $J \triangleq \{\boldsymbol{y} \in \mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \cap \mathcal{C} : \|\mathbf{x}^* - \boldsymbol{y}\|_2 \leq 1/2d^2\}$, the length of J is at least $\frac{1}{2d^2}$, and hence by Lemma 5.1, for every $\boldsymbol{y} \in J$ we have: $|p(\boldsymbol{y})| \geq |p(\mathbf{x}^*)|/2 \geq |p(\boldsymbol{a})|/2$. $\square$

Next, we wish to find a relatively large subcube of $\mathcal{C}$ of side length $\frac{1}{2d^2}$ for which all the valuations of p are large. Such a subcube may not exist. So instead, we prove the existence of a set of points in $\mathcal{C}$, of the same measure on which p can be guaranteed to have large values. We next define a measure of a set of points in a way that allows us to apply our inductive argument. This measure has the feature that when applied to the entire cube it becomes the standard integral.

Definition 5.4 (Measure)**.** *For any $k \in [n]$, and a fixed* affinity *point $\boldsymbol{a} = (a_{k+1}, \ldots, a_n) \in [-1,1]^{n-k}$, consider the k-dimensional* affine *cube $\mathcal{C}(\boldsymbol{a}) \triangleq [-1,1]^k \times \{a_{k+1}\} \times \cdots \times \{a_n\} \subseteq \mathcal{C}$. Let $J_k \subseteq \mathcal{C}(\boldsymbol{a})$. It's k-dimensional measure is defined as:*

$$|J_k| \triangleq \int_{\mathbf{x} \in J_k} dx_k \ldots dx_1.$$

[11] Here $\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}$ being *axis-parallel* is crucial, as t then varies only along $\boldsymbol{e}_i$, ensuring the degree of $p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}(t)$ is bounded by the individual degree of p.

Lemma 5.5 (Generalization of Lemma 5.1)**.** *Let $p\colon \mathcal{C} \to \mathbb{R}$ be a polynomial of individual degree at most d. For any $\boldsymbol{y} \in \mathcal{C}$ there exists a set of points $J \subset \mathcal{C}$, such that for every $\mathbf{x} \in J$, $|p(\mathbf{x})| \geq \frac{|p(\boldsymbol{y})|}{2^n}$, and*

$$|J| = \int_{\mathbf{x}=(x_1,\ldots,x_n)\in J} dx_n \ldots dx_1 \geq \frac{1}{(2d^2)^n}.$$

Proof. Fix some $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{C}$. Let $r \triangleq \frac{1}{2d^2}$. We build this set $J \subset [-1,1]^n$ inductively:

Base Case ($n = 1$)**:** Consider the line

$$\mathrm{L}_{\boldsymbol{y},\boldsymbol{e}_1} \triangleq \{\boldsymbol{y} + t\boldsymbol{e}_1 : t \in \mathbb{R}\}.$$

p restricted to $\mathrm{L}_{\boldsymbol{y},\boldsymbol{e}_1}$, i.e., $p_{\mathrm{L}_{\boldsymbol{y},\boldsymbol{e}_1}}(t) \triangleq p(\boldsymbol{y} + t\boldsymbol{e}_1)$ is a univariate polynomial in the variable t, of degree at most d. So, by Corollary 5.3, there exists a line segment $J_1 \subset \mathrm{L}_{\boldsymbol{y},\boldsymbol{e}_1} \cap \mathcal{C}$ such that $|p(\mathbf{x})| \geq \frac{|p(\boldsymbol{y})|}{2}$ for every $\mathbf{x} = (x_1, y_2, \ldots, y_n) \in J_1$, and $|J_1| = \int_{x_1 \in J_1} dx_1 \geq r$.

Induction Step ($n = k+1$)**:** Assume by *IH for* $n = k$ that there exists $J_k \subset [-1,1]^k$ such that, for every $\mathbf{x} = (x_1, \ldots, x_k, y_{k+1}, \ldots, y_n) \in J_k$, we have $|p(x_1, \ldots, x_k, y_{k+1}, \ldots, y_n)| \geq \frac{|p(\boldsymbol{y})|}{2^k}$, and

$$|J_k| = \int_{\mathbf{x}\in J_k} dx_k \ldots dx_1 \geq r^k.$$

Now, consider an arbitrary $\mathbf{x} \in J_k$ such that $\mathbf{x} = (x_1, \ldots, x_k, y_{k+1}, \ldots, y_n)$, and the line

$$\mathrm{L}_{(x_1,\ldots,x_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_{k+1}} \triangleq \{(x_1, \ldots, x_k, y_{k+1} + t, y_{k+2}, \ldots, y_n) : t \in \mathbb{R}\}.$$

Again, by Corollary 5.3, there exists $J_{k+1}^{(x_1,\ldots,x_k)} \subset \mathrm{L}_{(x_1,\ldots,x_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_{k+1}} \cap \mathcal{C}$, such that, for every

$$\boldsymbol{z} = (x_1, \ldots, x_k, z_{k+1}, y_{k+2}, \ldots, y_n) \in J_{k+1}^{(x_1,\ldots,x_k)}$$

we have,

$$|p(\boldsymbol{z})| = |p(x_1, \ldots, x_k, z_{k+1}, y_{k+2}, \ldots, y_n)| \geq \frac{|p(x_1, \ldots, x_k, y_{k+1}, \ldots, y_n)|}{2} \geq \frac{|p(\boldsymbol{y})|}{2^{k+1}},$$

and

$$\left|J_{k+1}^{(x_1,\ldots,x_k)}\right| = \int_{\boldsymbol{z}\in J_{k+1}^{(x_1,\ldots,x_k)}} dz_{k+1} \geq r. \tag{11}$$

Observe, for any $(x_1, \ldots, x_k) \neq (u_1, \ldots, u_k)$, $\mathrm{L}_{(x_1,\ldots,x_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_{k+1}} \cap \mathrm{L}_{(u_1,\ldots,u_k,y_{k+1},\ldots,y_n),\boldsymbol{e}_{k+1}} = \emptyset$. So, we may deduce $J_{k+1}^{(x_1,\ldots,x_k)} \cap J_{k+1}^{(u_1,\ldots,u_k)} = \emptyset$. So, for every $\mathbf{x} \in J_k$, the corresponding line segment along $\boldsymbol{e}_{k+1}$ is unique. Note that, for any $\mathbf{x} \in J_k$, from (11) we may infer:

$$\int_{\boldsymbol{z}\in J_{k+1}^{(x_1,\ldots,x_k)}} dz_{k+1} \geq \int_0^r dz_{k+1} = r. \tag{12}$$

So, all the line segments can be translated along $\boldsymbol{e}_{k+1}$, to cover the interval $[0, r]$. Figure 3 illustrates this idea, which allows us to construct higher dimensional cuboids. We define

$$J_{k+1} \triangleq \bigcup_{(x_1,\ldots,x_k,y_{k+1},\ldots,y_n)\in J_k} J_{k+1}^{(x_1,\ldots,x_k)}.$$

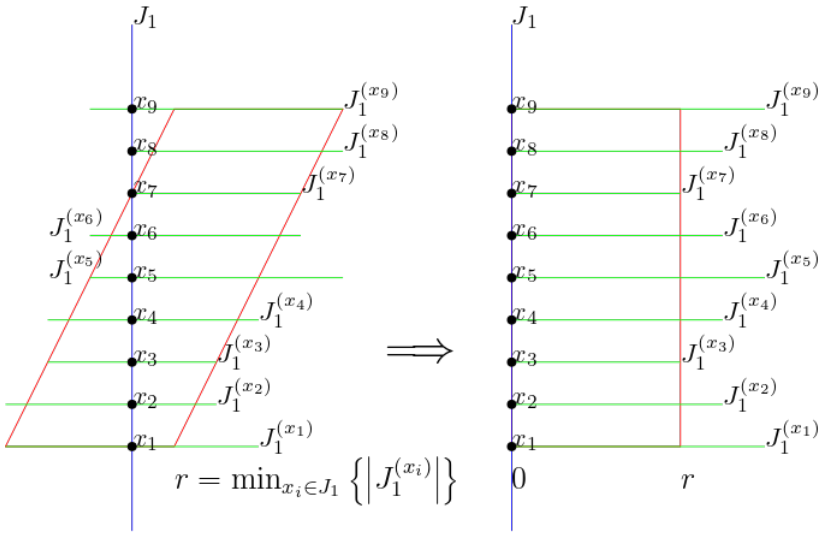

Figure 3: Translation for constructing J_2: For every $\mathbf{x}_i \in J_1$ (blue), the line segments $J_1^{(\mathbf{x}_i)}$ (green) are translated to have their left ends at 0, so that they all cover the rectangle (red) of height $|J_1| \geq r$, and width r. Thus, $|J_2| \geq r^2$.

Since, all the line segments $J_{k+1}^{(x_1,\ldots,x_k)}$ are unique, we have:

$$\begin{aligned}
|J_{k+1}| &\triangleq \int_{\mathbf{v}=(x_1,\ldots,x_{k+1},y_{k+2},\ldots,y_n)\in J_{k+1}} dx_{k+1}\ldots dx_1 \\
&= \int_{\mathbf{x}\in J_k}\int_{\boldsymbol{z}\in J_{k+1}^{(x_1,\ldots,x_k)}} dz_{k+1}dx_k\ldots dx_1 \\
&\geq \int_{\mathbf{x}\in J_k}\underbrace{\int_0^r dz_{k+1}}_{=r} dx_k\ldots dx_1 && \text{(From (12))} \\
&\geq r\underbrace{\int_{\mathbf{x}\in J_k} dx_k\ldots dx_1}_{=|J_k|\geq r^k} \geq r^{k+1}.
\end{aligned}$$

Thus, there exists $J = J_n \subset \mathcal{C}$, such that $|J| \geq r^n$, and for every $\mathbf{x} \in J$, $|p(\mathbf{x})| \geq \frac{|p(\boldsymbol{y})|}{2^n}$. □

The main theorem of this section now easily follows from Lemma 5.5.

Proof of Theorem 1.9. Let $\mathbf{x}^* \in \mathcal{C}$, such that $|p(\mathbf{x}^*)| = \|p\|_{\mathcal{C}_n,\infty}$. By Lemma 5.5, we have a $J \subset \mathcal{C}$, such that $|J| \geq \frac{1}{(2d^2)^n}$, and for every $\mathbf{x} \in J$, $|p(\mathbf{x})| \geq \frac{|p(\mathbf{x}^*)|}{2^n}$. Then,

$$\|p\|_{\mathcal{C}_n,1} = \int_{\mathcal{C}} |p(\mathbf{x})|d\mathbf{x} \geq \int_J |p(\mathbf{x})|d\mathbf{x} \geq \int_J \frac{|p(\mathbf{x}^*)|}{2^n}d\mathbf{x} = \frac{|p(\mathbf{x}^*)|}{2^n}|J| \geq \frac{\|p\|_{\mathcal{C}_n,\infty}}{(2d)^{2n}}.$$ □

5.1 Tightness of Theorem 1.9

We note that the lower bound of $\frac{1}{(2d)^{2n}}$ for $\frac{\|p\|_{\mathcal{C}_n,1}}{\|p\|_{\mathcal{C}_n,\infty}}$ is asymptotically tight, complementing Theorem 1.9:

Proposition 1.10. *There exists a global constant $c > 0$ such that for every odd d, there exists a family of individual degree-d polynomials $\{f_n\}_{n\in\mathbb{N}}$, where $f_n : \mathcal{C}_n \to \mathbb{R}$, such that $\|f_n\|_{\mathcal{C}_n,\infty} \geq c^n d^{2n}\|f_n\|_{\mathcal{C}_n,1}$.*

A simple observation towards this claim, may first be noted:

Observation 5.6. *There exists a family of a individual degree-d polynomials $\{h_n\}_{n\in\mathbb{N}}$, where $h_n : \mathcal{C}_n \to \mathbb{R}$, such that $\|h_n\|_{\mathcal{C}_n,\infty} \geq (d/2)^n \|h_n\|_{\mathcal{C}_n,1}$. Define $h_n(\boldsymbol{x}) \triangleq \prod_{i=1}^n x_i^d$, and observe $\|h_n\|_{\mathcal{C}_n,\infty} = 1$, while*

$$\|h_n\|_{\mathcal{C}_n,1} = \int_{\boldsymbol{x}\in\mathcal{C}_n} \prod_{i=1}^n |x_i^d| d\boldsymbol{x} = \prod_{i=1}^n \int_{-1}^1 |x_i^d| dx_i = \left(2\int_0^1 x^d dx\right)^n = \left(\frac{2}{d+1}\right)^n \leq \left(\frac{2}{d}\right)^n \|h\|_{\mathcal{C}_n,\infty}.$$

A tighter bound maybe shown using Legendre polynomials, as pointed out by [Hel18]:

Definition 5.7. *Legendre Polynomials are a family of orthogonal polynomials $\{P_n : [-1,1] \to [-1,1]\}_{n\in\mathbb{N}}$, indexed by their degree n, such that $P_0(x) = 1$, for all $x \in [-1,1]$, $P_n(1) = 1$, for all n, and for all $m \neq n$,*

$$\int_{-1}^1 P_n(x) P_m(x) dx = 0. \tag{13}$$

We note some useful properties of Legendre polynomials:

Fact 5.8. *Legendre Polynomials are even or odd, according as, their degree is even or odd, i.e.*

$$P_n(-x) = (-1)^n P_n(x). \tag{14}$$

Accordingly their derivatives, $P_n' : [-1,1] \to \mathbb{R}$, follow a similar rule:

$$P_n'(-x) = (-1)^{n+1} P_n'(x). \tag{15}$$

Their derivatives attain their respective extrema at the end points of the $[-1,1]$ interval:

$$\|P_n'\|_{[-1,1],\infty} = P_n'(1) = \frac{n(n+1)}{2}, \qquad \text{and} \qquad P_n'(-1) = (-1)^{n+1}\frac{n(n+1)}{2}. \tag{16}$$

For any $n \in \mathbb{N}$, as any polynomial $f(x)$ of degree $m < n$ can be written as a real linear combination of $\{P_i(x)\}_{i=0}^m$, i.e. $f(x) \equiv \sum_{i=0}^m c_i P_i(x)$ for some $\{c_i \in \mathbb{R}\}_{i=0}^m$, using (13) we get

$$\int_{-1}^1 P_n(x) f(x) dx = \sum_{i=0}^m c_i \underbrace{\int_{-1}^1 P_n(x) P_i(x) dx}_{=0,\because i\neq n} = 0. \tag{17}$$

Remark 5.9 ([Hel18]). *Let $m \in \mathbb{N}$. Consider the Legendre polynomial P_{m+1}, and define a degree-$(2m+1)$ polynomial $p : [0,1] \to \mathbb{R}$ as $p(x) \triangleq x(P_{m+1}'(2x-1))^2$. Then, $\|p\|_{[0,1],\infty} \geq (m+1)(m+2)\|p\|_{[0,1],1}$.*

Proof. From (16), we have $\max_{x\in[0,1]}\{|P_{m+1}'(2x-1)|\} = \|P_{m+1}'\|_{[-1,1],\infty} = P_{m+1}'(1)$. So, we observe:

- $p(x) \geq 0$ for every $x \in [0,1]$, and
- $\|p\|_{[0,1],\infty} = \max_{x\in[0,1]}\{x(P_{m+1}'(2x-1))^2\} = p(1) = (P_{m+1}'(1))^2$.

Since P_{m+1}' is a degree-m polynomial, for some degree-$(m-1)$ polynomial $q : \mathbb{R} \to \mathbb{R}$, we may write

$$P_{m+1}'(2x-1) = P_{m+1}'(1) + (x-1)q(2x-1). \tag{18}$$

$$\implies \|p\|_{[0,1],1} = \int_0^1 x(P_{m+1}'(1) + (x-1)q(2x-1))P_{m+1}'(2x-1)dx \qquad \text{(By (18))}$$

$$= P'_{m+1}(1) \underbrace{\int_0^1 x P'_{m+1}(2x-1)dx}_{\triangleq I_1} + \underbrace{\int_0^1 x(x-1)q(2x-1)P'_{m+1}(2x-1)dx}_{\triangleq I_2}.$$

Using integration by parts on the second integral, we get

$$\begin{aligned} I_2 &= \int_0^1 x(x-1)q(2x-1)P'_{m+1}(2x-1)dx \\ &= \Big[\underbrace{x(x-1)q(2x-1)\int P'_{m+1}(2x-1)dx}_{=(x(x-1)q(2x-1)P_{m+1}(2x-1))/2}\Big]_0^1 - \int_0^1 P_{m+1}(2x-1)\underbrace{\frac{d(x(x-1)q(2x-1))}{dx}}_{\triangleq r(2x-1),\ deg(r)\le m}dx \\ &= -\int_0^1 P_{m+1}(2x-1)r(2x-1)dx = -\frac{1}{2}\int_{-1}^1 P_{m+1}(y)r(y)dy = 0. \qquad \text{(By (17))} \end{aligned}$$

Using integration by parts, this time on the first integral, we get

$$\begin{aligned} I_1 &= \int_0^1 x P'_{m+1}(2x-1)dx = \left[x\int P'_{m+1}(2x-1)dx\right]_0^1 - \int_0^1 \left(\int P'_{m+1}(2x-1)dx\right)dx \\ &= \left[\frac{x}{2}P_{m+1}(2x-1)\right]_0^1 - \frac{1}{2}\int_0^1 P_{m+1}(2x-1)dx = \frac{P_{m+1}(1)}{2} - \frac{1}{4}\underbrace{\int_{-1}^1 P_0(y)P_{m+1}(y)dy}_{=0,\ by\ (13)}. \end{aligned}$$

$$\implies \|p\|_{[0,1],1} = \frac{P'_{m+1}(1)P_{m+1}(1)}{2} = \frac{P'_{m+1}(1)}{2}. \qquad (P_m(1)=1, \text{ for all } m\in\mathbb{N})$$

$$\implies \frac{\|p\|_{[0,1],1}}{\|p\|_{[0,1],\infty}} = \frac{P'_{m+1}(1)}{2(P'_{m+1}(1))^2} = \frac{1}{2P'_{m+1}(1)} = \frac{1}{(m+1)(m+2)}. \qquad \text{(By (16))}$$

□

Using this, we can now prove the main theorem of this sub-section:

Proof of Proposition 1.10. Let $d = 2m+1$ for some $m \in \mathbb{N}$. Define a degree-$(2m+1)$ polynomial $f : [-1,1] \to \mathbb{R}$, as $f(x) = p\left(\frac{x+1}{2}\right)$, where $p : [0,1] \to \mathbb{R}$ is as defined in Remark 5.9. Then:

$$\begin{aligned} &\|f\|_{[-1,1],\infty} = \|p\|_{[0,1],\infty}, \text{ and} \\ &\|f\|_{[-1,1],1} = \int_{-1}^1 |f(x)|dx = \int_{-1}^1 \left|p\left(\frac{x+1}{2}\right)\right|dx = 2\int_0^1 |p(y)|dy = 2\|p\|_{[0,1],1}. \end{aligned}$$

From Remark 5.9, we have

$$\|f\|_{[-1,1],\infty} = \|p\|_{[0,1],\infty} = (m+1)(m+2)\|p\|_{[0,1],1} = \frac{(m+1)(m+2)}{2}\|f\|_{[-1,1],1}.$$

For any $n \in \mathbb{N}$, define $f_n(\boldsymbol{x}) \triangleq \prod_{i=1}^n f(x_i)$. Observe $\|f_n\|_{\mathcal{C}_n,\infty} = (\|f\|_{[-1,1],\infty})^n$, while

$$\|f_n\|_{\mathcal{C}_n,1} = \int_{\boldsymbol{x}\in\mathcal{C}_n}\prod_{i=1}^n |f(x_i)|d\boldsymbol{x} = \prod_{i=1}^n \int_{-1}^1 |f(x_i)|dx_i = (\|f\|_{[-1,1],1})^n.$$

$$\implies \frac{\|f_n\|_{\mathcal{C}_n,\infty}}{\|f_n\|_{\mathcal{C}_n,1}} = \left(\frac{\|f\|_{[-1,1],\infty}}{\|f\|_{[-1,1],1}}\right)^n = \left(\frac{(m+1)(m+2)}{2}\right)^n \ge \left(\frac{1}{8}\right)^n d^{2n}. \qquad \square$$

6 Robust multivariate regression algorithm

In this section, we show how a modification of our algorithm that avoid a run-time dependence on $(\|p\|_{\mathcal{C}_n,\infty}, \sigma, N)$ and thus prove the following theorem.

Theorem 6.1. *[Generalized form of Theorem 1.3] Let $\varepsilon \in (0, 1/2]$,$\delta \in (0, \varepsilon]$, $\sigma \geq 0$, and $\rho < 1/2$. There is an algorithm that solves the Robust Multivariate Polynomial Regression Problem with approximation factor $C = 2 + \varepsilon$. The output of the algorithm is a polynomial $\widehat{p}\colon \mathbb{R}^n \to \mathbb{R}$ of degree at most d in each variable, such that with probability (over the random input samples) at least $1 - \delta$, $\widehat{p}$ satisfies*

$$|p(\mathbf{x}) - \widehat{p}(\mathbf{x})| \leq (2 + \varepsilon)\sigma, \qquad \text{for all } \mathbf{x} \in \mathcal{C}\,.$$

It uses $M = O_{n,\rho}\left((d^{2n+1}/\varepsilon)^n \log(d/\delta)\right)$ samples drawn from the multidimensional Chebyshev distribution, or $M = O_{n,\rho}\left((d^{2n+1}/\varepsilon)^{2n} \log(d/\delta)\right)$ if the sampled are drawn from the uniform measure. Its run-time is $\mathsf{poly}(M, \log_\varepsilon(1 - 2\rho))$.

The first step is to generalize KKP's ℓ_1 regression. Similar to their regression on averages over Chebyshev intervals, we do regression on averages over Chebyshev cells.

Definition 6.2 (ℓ_1 Minimizer)**.** *Let m be a large enough integer. Given a set of M samples $S = \{(\mathbf{x}_i, y_i)\}$, for every $\boldsymbol{j} \in [m]^n$, let $S_{\boldsymbol{j}} \triangleq \{\beta \in [M] : \mathbf{x}_\beta \in \mathcal{C}_{\boldsymbol{j}}\}$. The ℓ_1 minimizer of S with respect to the (m, n)-Chebyshev partition, is the individual degree-d polynomial:*

$$\widehat{p}_{\ell_1} \triangleq \arg\min_{f \in \mathcal{P}_d} \sum_{\boldsymbol{j} \in [m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\beta \in S_{\boldsymbol{j}}} |f(\mathbf{x}_\beta) - y_\beta|. \tag{19}$$

We show that on a set of α-good samples, the ℓ_1 regression outputs an individual degree-d polynomial with $\mathsf{poly}(d^n)$ error in ℓ_∞.

Theorem 6.3 (ℓ_∞ error bound for the ℓ_1 minimizer)**.** *Let $\alpha < 1/2$ be constant , $\varepsilon \leq (1 - 2\alpha)/2$, and for some constant $c > 1$, $m \geq (cd)^{2n+1}/\varepsilon$. Given a set S of samples that is α-good with respect to the (m, n)-Chebyshev partition, the ℓ_1 minimizer $\widehat{p}_{\ell_1}$ from (19) satisfies*

$$\|p - \widehat{p}_{\ell_1}\|_{\mathcal{C}_n,\infty} = O_\alpha((8d^2)^n \sigma).$$

This proof is provided in Appendix C. In Algorithm 3, we use the same idea as in KKP: starting with $\widehat{p}_{\ell_1}$, the result of an ℓ_1 regression, we then iteratively refine the estimate, improving the ℓ_∞ error in each step.

Algorithm 3: Median Based Recovery, with ℓ_1 regression

Given : A set of samples $S = \{\mathbf{x}_i, y_i\}_{i=1}^M$, of which a ρ fraction may be outliers.

$\widehat{p}^{(0)} \leftarrow$ result of ℓ_1 regression: $\widehat{p}_{\ell_1}$;

$N_3 \leftarrow O\left(n \log_{1/\varepsilon} d + \log_\varepsilon(1 - 2\rho)\right)$;

for $i \in \{0, \ldots, N_3 - 1\}$ **do**

  $\widehat{p}^{(i+1)} \leftarrow$ REFINE$(S, \widehat{p}^{(i)})$;

Return $\widehat{p}^{(N_3)}$.

As a result, the number of iterations here depends only on (d, n), and improves as ε gets smaller. (In fact, it is linear in n, and logarithmic in d, and $1/\varepsilon$).

Here, as in Section 3, we prove that with enough samples, the set of samples is good with high probability, and separately we show that for any α-good set of samples Algorithm 3 recovers p as required.

Theorem 6.4 (Absolute ℓ_∞ error bound)**.** *Let $\varepsilon, \alpha < 1/2$. Let the set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^M$ of samples be α-good for the (m,n)-Chebyshev partition where $m \geq (cd)^{2n+1}/\varepsilon$, for some large enough constant $c > 1$. Then the median recovery Algorithm 3, in $N_3 = O(n \log_{1/\varepsilon} d + \log_\varepsilon(1-2\alpha))$ iterations, returns an individual degree-d polynomial $\widehat{p}$, such that $\|p - \widehat{p}\|_{\mathcal{C}_n,\infty} \leq (2+\varepsilon)\sigma$.*

Proof. Using the same notations as in the proof of Theorem 3.4, we conclude

$$\|p - \widehat{p}^{(t)}\|_{\mathcal{C}_n,\infty} \leq (2+6\varepsilon)\sigma + \varepsilon^t \|e_0\|_{\mathcal{C}_n,\infty}.$$

In Algorithm 3, $\widehat{p}^{(0)}$ is set to be the result of ℓ_1 regression, giving us $e_0 = p - \widehat{p}_{\ell_1}$. Since S is assumed to be α-good for the (m,n)-Chebyshev partition, where $m = \Omega((cd)^{2n+1}/\varepsilon)$, for some absolute constant $c > 1$, applying Theorem 6.3, we get $\|e_0\|_{\mathcal{C}_n,\infty} \leq (8d^2)^n O_\alpha(\sigma)$. Thus,

$$\|p - \widehat{p}^{(t)}\|_{\mathcal{C}_n,\infty} \leq (2+6\varepsilon)\sigma + \varepsilon^t (2\sqrt{2}d)^{2n} \frac{4\sigma}{1-2\alpha}.$$

So, in $N_3 \geq \log_{1/\varepsilon} \frac{(2\sqrt{2}d)^{2n}}{1-2\alpha}$ iterations, followed by a further rescaling of ε to $\varepsilon/10$, we get the desired error bound. □

Now we are ready to prove the main theorem of this section:

Proof of Theorem 6.1. It follows the same arguments as the proof of Theorem 3.1, with Theorem 3.4 being replaced by Theorem 6.4, according to which: for Algorithm 3, $m = (cd)^{2n+1}/\varepsilon$ suffices, for some absolute constant $c > 1$. This gives us the Chebyshev, and Uniform sample complexities of $M_C = \frac{1}{(1-2\rho)^2}\left(\frac{cd^{2n+1}}{\varepsilon}\right)^n \log\frac{d}{\varepsilon\delta}$, and $M_U = \frac{1}{(1-2\rho)^2}\left(\frac{cd^{2n+1}}{\varepsilon}\right)^{2n} \log\frac{d}{\varepsilon\delta}$, respectively, for some constant $c > 1$.

The value of α that we set to deal with $\rho(< 0.5)$ fraction of outliers remains the same: $\alpha = \frac{2\rho+1}{4} < \frac{1}{2}$. So, the run-time is that of solving $N_3 = O(n \log_{1/\varepsilon} d + \log_\varepsilon(1-2\rho))$ linear programs with $O(d^n)$ variables, and M constraints. □

7 Sample Complexity Lower Bounds

7.1 Sample complexity lower bound against uniform sampling

In this section, we prove Theorem 1.4, restated here for convenience:

Theorem 1.4. *For any approximation factor $C > 1$, there exists $c = c(C) > 0$ such that any algorithm, that solves the Robust Multivariate Polynomial Regression Problem, requires at least $(cd)^{2n}$ samples drawn from the uniform measure to succeed with probability more than $2/3$. This holds for any outlier probability ρ.*

Before we prove Theorem 1.4, we formally note the following definitions:

Definition 7.1 (Chebyshev Polynomials)**.** *Chebyshev polynomials of the first kind are degree-d polynomials $T_d : \mathbb{R} \to \mathbb{R}$, that follow the recurrence relation:*

$$T_0(x) = 1, \qquad T_1(x) = x, \qquad T_{d+1}(x) = 2xT_d(x) - T_{d-1}(x).$$

$$T_d(x) \triangleq \begin{cases} \cos(d \arccos(x)), & \text{if } |x| \leq 1, \\ \cosh(d \operatorname{arcosh}(x)), & \text{if } x \geq 1, \\ (-1)^d \cosh(d \operatorname{arcosh}(-x)), & \text{if } x \leq -1, \end{cases} \quad \textit{is their explicit trigonometric formulation.}$$

Definition 7.2 (Chebyshev Extremas)**.** *For any $d \in \mathbb{Z}_{>0}$, the d Chebyshev extremas $\in [-1,1]$ given by*

$$x_k \triangleq \cos\left(\frac{k}{d}\pi\right), k \in [d]$$

are the extremas of T_d, the degree-d Chebyshev polynomial of the first kind, i.e., $T_d(x_k) \in \{\pm 1\}, \forall k \in [d]$.

Proof of Theorem 1.4. For $n = 1$, KKP showed that the polynomial $T_d\left(x_1 + \frac{\alpha}{d^2}\right)$ (for some constant α dependent on C) and the identically 0 polynomial are indistinguishable with high probability. We build on their result. Let $\alpha = 4\sqrt{C}$ and consider the following multivariate polynomial, with degree d in each variable:

$$f(\mathbf{x}) \triangleq \frac{\prod_{i=1}^{n} T_d\left(x_i + \frac{\alpha}{d^2}\right)}{\left(T_d\left(1 + \frac{\alpha}{d^2}\right)\right)^{n-1}}.$$

Let the target polynomial p to be learned be f with probability $1/2$ and otherwise, $g \equiv 0$. Then,

(i) For every $\mathbf{x} \in \mathcal{C}_n \setminus \left[1 - \frac{\alpha}{d^2}, 1\right]^n$ there exists at least one index $i \in [n]$ such that $x_i + \frac{\alpha}{d^2} \le 1$, implying $\left|T_d\left(x_i + \frac{\alpha}{d^2}\right)\right| \le 1$. For all other $j \neq i \in [n]$, we have $\left|\frac{T_d\left(x_j + \frac{\alpha}{d^2}\right)}{T_d\left(1 + \frac{\alpha}{d^2}\right)}\right| \le 1$, since $T_d(1) = 1$, and $T_d(x)$ is monotonically increasing on the region $x \ge 1$. Hence, $|f(\mathbf{x}) - g(\mathbf{x})| = |f(\mathbf{x})| \le 1$. We thus say that f and g are indistinguishable on the region $\mathcal{C}_n \setminus \left[1 - \frac{\alpha}{d^2}, 1\right]^n$ for noise level $\sigma = 1$. We assume $d > \sqrt{\alpha/2}$, as otherwise, the theorem follows trivially for $c = \sqrt{2/\alpha}$. Thus, the probability that $\mathbf{x}$ lands in the *distinguishable* region is $= \frac{V_n\left([1-\alpha/d^2, 1]^n\right)}{V_n(\mathcal{C}_n)} = \frac{1}{2^n}\left(\frac{\alpha}{d^2}\right)^n$.

(ii) For all $d > 1$ we have $\|f - g\|_{\mathcal{C}_n,\infty} > 2C$. This follows from

$$\|f - g\|_{\mathcal{C}_n,\infty} \ge |f(\mathbf{1}) - g(\mathbf{1})| = \left|T_d\left(1 + \frac{\alpha}{d^2}\right)\right| > \frac{\alpha^2}{8} = 2C.$$

For all $\alpha > 0$, the second inequality can be verified for $d = 2$ (where $T_2(x) = 2x^2 - 1$), and follows for all $d > 2$ since $T_d\left(1 + \frac{\alpha}{d^2}\right) = \cosh\left(d \operatorname{arcosh}\left(1 + \frac{\alpha}{d^2}\right)\right)$ is increasing in d, as KKP noted.

Then, for $M < (cd)^{2n}$ independent samples uniformly generated over $\mathcal{C}$, the event that all of them land in the indistinguishable region, happens with probability:

$$\begin{aligned}\Pr\left[\forall i \in [M].\, \mathbf{x}_i \in \mathcal{C}_n \setminus [1 - \alpha/d^2, 1]^n\right] &\ge 1 - M \cdot \Pr\left[\mathbf{x} \notin \mathcal{C}_n \setminus [1 - \alpha/d^2, 1]^n\right] \\ &= 1 - M\left(\frac{\alpha}{2d^2}\right)^n > \frac{2}{3}.\end{aligned}$$

The first inequality is by union bound over the M samples and the last inequality follows from $M < \frac{1}{3}\left(\frac{2d^2}{\alpha}\right)^n$ (e.g., for $c = (36C)^{-1/4}$). Consider the following adversarial strategy: the adversary chooses $p' = f$ with probability $1/2$, and otherwise $p' = g$. For samples in the indistinguishable region, it outputs p', and otherwise p. Now, with probability $> 2/3$, all the samples are from the indistinguishable region. So, with this probability, the algorithm observes only values of p', which are independent of the target function p, and the best it can do is to guess the function with probability $1/2$. Thus, it fails with probability $> 2/3 \cdot 1/2 = 1/3$. □

Remark 7.3. *We can get a lower bound of $(cd/n)^{2n}$ (with the same value of c as in Theorem 1.4) for the sample complexity of learning* total degree-d *polynomials in our setting, by simply applying the above argument with the individual degree being $\lfloor d/n \rfloor$. One can get a slightly improved bound of $(cd/\sqrt{n})^{2n}$ for this problem by using the polynomial $T_d\left(\frac{\alpha}{d^2} + \frac{1}{n}\sum x_i\right)$, with the same values of α and c as above. Here, the indistinguishable region becomes $\mathcal{C}_n \setminus [1 - n\alpha/d^2, 1]^n$, and so, the analysis slightly changes.*

7.2 Distribution-free sample complexity lower bound

In this section, we prove Theorem 1.5. We start with some auxiliary lemmas. We begin by generalizing [KKP17, Lemma 5.2], which we state for reference:

Lemma 7.4 ([KKP17], Lemma 5.2)**.** *Let $d \geq 1$. For any point $a \in [-1,1]$, there exists a degree-d polynomial $q_a : \mathbb{R} \to \mathbb{R}$, such that $q_a(a) = 1 = \|q_a\|_{[-1,1],\infty}$, and for every $x \in [-1,1]$,*

$$|q_a(x)| \leq \frac{2}{d|x-a|}.$$

Next, we generalize [KKP17, Lemma 5.3], a variant of which we state for reference, and then briefly describe their argument:

Lemma 7.5 ([KKP17], Lemma 5.3)**.** *For any $d, \alpha > 0$, let $m = \lfloor d\alpha/2 \rfloor$. Define a set of nodes $b_j \triangleq -1 + \frac{2j}{m}, j \in [m]$. For any $S \subseteq [m]$, consider the set of degree-d polynomials*

$$g_S(y) \triangleq \sum_{j \in S} q_{b_j}(y),$$

where $q_{b_j} : [-1,1] \to \mathbb{R}$ is the degree-d polynomial guaranteed from [KKP17, Lemma 5.2]. For any $y \in [-1,1]$, define $k_y \triangleq \arg\min_{j \in [m]} |b_j - y|$, i.e. the index of the node b_j closest to y. Then, for any $j \in [m]$, if $j \neq k_y$ then

$$|g_{\{j\}}(y)| \leq \alpha.$$

Now, for the (univariate) lower bound construction (in short): the adversary

- picks a random $S \subseteq [m]$, and an arbitrary j^* out of the set of outlier full nodes (i.e. those nodes for which all the sample points, that these nodes are the nearest nodes, are outliers),
- defines $S' \triangleq S \Delta \{j^*\}$ (i.e. $S' = S \cup \{j^*\}$, if $\{j^*\} \notin S$, and $S' = S \setminus \{j^*\}$, if $\{j^*\} \in S$) so that $S \Delta S' = \{j^*\}$, and
- replies with either f_S or $f_{S'}$ with probability $1/2$.

Note, for every sample x we have $|f_S(x) - f_{S'}(x)| = |f_{\{j^*\}}(x)|$. Either

- $k_x = j^*$ i.e. $x \in L_{j^*}$ is outlier (in which case $|f_S(x) - f_{S'}(x)|$ is arbitrary), or
- $k_x \neq j^*$, implying $|x - b_{j^*}| \geq 1/m$, and hence, by Lemma 7.5

$$|f_S(x) - f_{S'}(x)| = |f_{\{j^*\}}(x)| = |q_{b_j^*}(x)| \leq \frac{2}{d|x - b_{j^*}|} \leq \alpha.$$

The last inequality follows for $m = d\alpha/2$. We now generalize Lemma 7.4.

Lemma 7.6. *Let $d \geq 1$. For any $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathcal{C}$, there exists a polynomial $p_{\boldsymbol{b}} : \mathcal{C} \to \mathbb{R}$ with individual degree at most d, such that $|p_{\boldsymbol{b}}(\boldsymbol{b})| = 1$, and for all $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{C}$,*

$$|p_{\boldsymbol{b}}(\boldsymbol{x})| \leq \frac{2}{d \max_{i \in [n]} |x_i - b_i|}.$$

Proof of Lemma 7.6. Fix some $\boldsymbol{b} \in \mathcal{C}$. Consider the individual degree at most d polynomial $p_{\boldsymbol{b}} : \mathcal{C} \to \mathbb{R}$ defined as:

$$p_{\boldsymbol{b}}(\boldsymbol{x}) \triangleq \prod_{i=1}^{n} q_{b_i}(x_i). \tag{20}$$

Where q_{b_i} are the polynomials defined in Lemma 7.4. Since for all $i \in [n]$, $|q_{b_i}(b_i)| = 1$, we have,

$$|p_{\boldsymbol{b}}(\boldsymbol{b})| = \prod_{i=1}^{n} |q_{b_i}(b_i)| = 1.$$

We next show that $|p_{\boldsymbol{b}}(x)|$ cannot be too large for all $\boldsymbol{x} \in \mathcal{C}$. Let i' be the index that achieves the maximum of $|x_i - b_i|$. For every $i \neq i'$ we use the bound $|q_{b_i}(x_i)| \leq 1$, while for i' we use the bound $|q_{b_i}(x_i)| \leq \frac{2}{d|x_i - b_i|}$, with both the bounds being from Lemma 7.4. We get,

$$|p_{\boldsymbol{b}}(\boldsymbol{x})| = \prod_{i=1}^{n} |q_{b_i}(b_i)| \leq \frac{2}{d \max_{i \in [n]} |x_i - b_i|}. \qquad \square$$

In order to generalize Lemma 7.5, let us define *nodes*, and the notion of *closest* nodes in the cube $\mathcal{C}$:

Definition 7.7. *For any $m > 0$, consider the set of points $b_j \triangleq -1 + \frac{2j}{m}, j \in [m]$, that equi-partition $[-1, 1]$ along each axes. For any $\boldsymbol{j} = (j_1, \ldots, j_n) \in [m]^n$, the corresponding* node *is $\boldsymbol{b_j} \triangleq (b_{j_1}, \ldots, b_{j_n})$.*

For any $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{C}$, the closest (in ℓ_1) node to $\mathbf{x}$ is $\boldsymbol{b}_{\boldsymbol{k}(\mathbf{x})}$, where $\boldsymbol{k}(\mathbf{x}) \triangleq (k_1, \ldots, k_n)$, and $k_i \triangleq \arg\min_{j \in [m]} |b_j - x_i|$, for all $i \in [n]$.

Lemma 7.8. *For any $d, \alpha > 0$, let $m = \lfloor d\alpha/2 \rfloor$. Consider the set of m^n* nodes *$\{\boldsymbol{b_j}, \boldsymbol{j} \in [m]^n\}$, and the notion of* closest nodes*, as discussed in Definition 7.7. For any subset of nodes, $\boldsymbol{S} \subseteq [m]^n$, define*

$$f_{\boldsymbol{S}}(\boldsymbol{x}) \triangleq \sum_{\boldsymbol{j} \in \boldsymbol{S}} p_{\boldsymbol{b_j}}(\boldsymbol{x}),$$

where $\boldsymbol{b_j} = (b_{j_1}, \ldots, b_{j_n})$, for $\boldsymbol{j} = (j_1, \ldots, j_n)$, and $p_{\boldsymbol{b_j}}$'s are individual degree-d polynomials from Lemma 7.6. For any $\mathbf{x} \in \mathcal{C}$, let the closest in ℓ_1 node to $\mathbf{x}$ be $\boldsymbol{b}_{\boldsymbol{k}(\mathbf{x})}$. Then, for any $\boldsymbol{j} \neq \boldsymbol{k}(\mathbf{x})$, $|f_{\{\boldsymbol{j}\}}(\mathbf{x})| \leq \alpha$.

Proof of Lemma 7.8. Fix $\boldsymbol{x} \in \mathcal{C}$, and $\boldsymbol{j} \neq \boldsymbol{k}(\mathbf{x})$. Then

$$|f_{\{\boldsymbol{j}\}}(\boldsymbol{x})| = |p_{\boldsymbol{b_j}}(\boldsymbol{x})| \leq \frac{2}{d \max_i |x_i - b_{j_i}|} \leq \alpha, \tag{21}$$

The first inequality is by Lemma 7.6 and the second holds for $\max_{i \in [n]} |x_i - b_{j_i}| \geq 1/m$. Indeed, since $\boldsymbol{j} \neq \boldsymbol{k}(\mathbf{x})$, there exists an index $\tau \in [n]$ for which $j_\tau \neq k_\tau$, and x_τ is closer to b_{k_τ} rather than b_{j_τ}. So, $|x_\tau - b_{j_\tau}| \geq |b_{k_\tau} - b_{j_\tau}|/2 \geq 1/m$. $\square$

Finally we prove the lower bound for polynomials of individual degrees at most d (Theorem 1.5, restated below).

Theorem 1.5. *For any approximation factor $C > 1$, and any outlier probability $\rho > 0$, there exists $c = c(C, \rho) > 0$ such that any algorithm, that solves the Robust Multivariate Polynomial Regression Problem, requires at least $(cd)^n \log d$ samples drawn from any measure over $[-1, 1]^n$ to succeed with probability more than $2/3$.*

Proof. We will prove a stronger bound, showing that $M < (cd)^n n \log d$ samples are not enough to succeed with constant probability. We may assume $d > 12C$ as otherwise the lower bound is trivial. Let $\sigma = \frac{1}{3C}$, $m = \lfloor \frac{d}{6C} \rfloor$, and assume $M < (cd)^n n \log d$.

Consider the set of nodes $\boldsymbol{b}_{\boldsymbol{j}}$ for all $\boldsymbol{j} \in [m]^n$ as defined in Definition 7.7. For every $\boldsymbol{j} = (j_1, \ldots, j_n) \in [m]^n$, let $L_{\boldsymbol{j}}$ be the set of samples for which the nearest node is $\boldsymbol{b}_{\boldsymbol{j}}$ in ℓ_1 distance. Formally, recall that for all $\mathbf{x} \in \mathcal{C}$, $\boldsymbol{k}(\mathbf{x})_i \triangleq \arg\min_j |b_j - x_i|$ where the minimization is over the one dimensional nodes $\{b_j\}_{j \in [m]}$, then

$$L_{\boldsymbol{j}} \triangleq \{\beta \in [M] : \boldsymbol{k}(\mathbf{x}_\beta) = \boldsymbol{j}\}.$$

We say that $L_{\boldsymbol{j}}$ is *outlier-full* if all the samples $(\mathbf{x}_\beta, y_\beta) \in L_{\boldsymbol{j}}$ are outliers. We say that $L_{\boldsymbol{j}}$ is *small*, if $|L_{\boldsymbol{j}}| \leq \frac{2M}{m^n}$. So, a *small* $L_{\boldsymbol{j}}$ is *outlier-full* with probability

$$\rho^{|L_{\boldsymbol{j}}|} \geq \rho^{2(cd/m)^n n \log d} \geq d^{-o(n)}.$$

The last inequality is by $m \geq \frac{d}{12C}$ for $d \geq 12C$ and holds for large enough $c = c(C, \rho)$. Since there are at least $m^n/2$ *small* $L_{\boldsymbol{j}}$'s, the probability that none of these *small* $L_{\boldsymbol{j}}$'s is *outlier-full* is at most

$$(1 - d^{-o(n)})^{m^n/2} \leq e^{-m^n d^{-o(n)}/2} < 1/3.$$

The first inequality is by $1 - x \leq e^{-x}$ for all x and the last inequality holds for $d > 12C$. So, the probability that at least one of the *small* $L_{\boldsymbol{j}}$'s is *outlier-full* is $\geq 2/3$.

Let $f_{\boldsymbol{S}}$ be the true polynomial to be learned, defined as in Lemma 7.8, for $\boldsymbol{S}$ chosen uniformly from $\mathcal{P}([m]^n)$. Consider the following adversarial strategy: given the samples and the outliers' locations, suppose there exists an *outlier-full* region, then the adversary picks an arbitrary $\boldsymbol{j}^*$ such that $L_{\boldsymbol{j}^*}$ is *outlier-full*. Let $\boldsymbol{S}' \triangleq \boldsymbol{S} \Delta \{\boldsymbol{j}^*\}$, the adversary chooses $p' = f_{\boldsymbol{S}}$ or $p' = f_{\boldsymbol{S}'}$ equiprobably, and replies $(\mathbf{x}_\beta, p'(\mathbf{x}_\beta))$, for all $\beta \in [M]$. Since $\boldsymbol{S} \Delta \boldsymbol{S}' = \{\boldsymbol{j}^*\}$, we have for all $\mathbf{x} \in \mathcal{C}$, $|f_{\boldsymbol{S}}(\mathbf{x}) - f_{\boldsymbol{S}'}(\mathbf{x})| = |f_{\{\boldsymbol{j}^*\}}(\mathbf{x})|$. For any sample $\beta \in [M]$, if this sample is not an outlier, then $\beta \notin L_{\boldsymbol{j}^*}$ i.e. $\boldsymbol{k}(\mathbf{x}_\beta) \neq \boldsymbol{j}^*$, and by Lemma 7.8, we have $|f_{\boldsymbol{S}}(\mathbf{x}) - f_{\boldsymbol{S}'}(\mathbf{x})| = |f_{\{\boldsymbol{j}^*\}}(\mathbf{x})| \leq \sigma$. Since $\boldsymbol{j}^*$ is chosen based on the locations of outliers, but independent of $\boldsymbol{S}$, the distributions of $\boldsymbol{S}'$ and $\boldsymbol{S}$ are the same. This makes $f_{\boldsymbol{S}}$ *indistinguishable* from $f_{\boldsymbol{S}'}$. We finally note,

$$\|f_{\boldsymbol{S}} - f_{\boldsymbol{S}'}\|_{\mathcal{C}_n,\infty} = \|f_{\{\boldsymbol{j}^*\}}\|_{\mathcal{C}_n,\infty} = \|p_{\boldsymbol{b}_{\boldsymbol{j}^*}}\|_{\mathcal{C}_n,\infty} \geq |p_{\boldsymbol{b}_{\boldsymbol{j}^*}}(\boldsymbol{b}_{\boldsymbol{j}^*})| = 1 > 2C\sigma,$$

so that, any output of the algorithm at least $C\sigma$ far in ℓ_∞ from either $f_{\boldsymbol{S}}$, or $f_{\boldsymbol{S}'}$. With probability at least $2/3$ there exists a *small outlier-full* $L_{\boldsymbol{j}^*}$, in this case, the best the algorithm can do is to guess and so the failure probability becomes at least $2/3 \cdot 1/2 = 1/3$. □

References

[AK03] Sanjeev Arora and Subhash Khot. Fitting algebraic curves to noisy data. *Journal of Computer and System Sciences*, 67(2):325–340, 2003. Special Issue on STOC 2002. 2, 6

[DDL18] Hadassa Daltrophe, Shlomi Dolev, and Zvi Lotker. Big data interpolation using functional representation. *Acta Informatica*, 55:213–225, 2018. 6, 11

[DiB02] Emmanuele DiBenedetto. *Real Analysis*. Birkhäuser, 2002. 13

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019. 6

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019. 6

[GZ16] V. Guruswami and D. Zuckerman. Robust Fourier and Polynomial Curve Fitting. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 751–759, Los Alamitos, CA, USA, oct 2016. IEEE Computer Society. 2, 6

[Hel18] Helmut. Norms on $\mathcal{P}_N$ Vector Space of Polynomials up to Order N. Mathematics Stack Exchange, 2018. https://math.stackexchange.com/q/2693954 (version: 2018-03-19). 23

[KKM18] Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1420–1430. PMLR, 2018. 6

[KKP17] Daniel Kane, Sushrut Karmalkar, and Eric Price. Robust Polynomial Regression up to the Information Theoretic Limit. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 391–402, 2017. 2, 4, 6, 7, 11, 12, 15, 28, 33

[Mar90] Andrey Andreyevich Markov. On a question by D. I. Mendeleev. *Zap. Imp. Akad. Nauk. St. Petersburg*, 62:1–24, 1890. https://history-of-approximation-theory.com/fpapers/markov4.pdf. 10

[Nev79] Paul G Nevai. Bernstein's inequality in lp for 0< p< 1. *Journal of Approximation Theory*, 27(3):239–243, 1979. 7

[PSBR20] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):601–627, 2020. 6

[Rud76] Walter Rudin. *Principles of Mathematical Analysis*. McGraw Hill, 1976. 20

[Wil74] Don R Wilhelmsen. A Markov Inequality in Several Dimensions. *Journal of Approximation Theory*, 11(3):216–220, 1974. 5, 36

[Wol06] John Wolberg. *Data analysis using the method of least squares: extracting the most information from experiments*. Springer Science & Business Media, 2006. 2

[Zie11] Achim Zielesny. *From curve fitting to machine learning*, volume 18. Springer, 2011. 2

A Exponential lower bound for linear functions

Theorem A.1. *For any constant approximation factor $C > 1$, and any $\sigma < \frac{1}{2C}$, given $M = e^{o(n\sigma^2)}$ samples, drawn from any product distribution with mean 0, no algorithm can solve the Robust Multivariate Polynomial Regression Problem with failure probability $\delta < 1/4$, for any ρ (even for $\rho = 0$, i.e., even without outliers).*

In particular, for constant noise level σ, any algorithm requires, $e^{\Omega(n)}$ many samples to succeed with probability at least $3/4$.

Proof of Theorem A.1. Consider the two linear functions: $g \equiv 0$,and $h(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$. On one hand, we have $\max_{\mathbf{x}\in[-1,1]^n} |g(\mathbf{x}) - h(\mathbf{x})| \geq |g(\mathbf{1}) - h(\mathbf{1})| \geq 1$. On the other hand, for many $\mathbf{x} \in \mathcal{C}$, $|g(\mathbf{x}) - h(\mathbf{x})| \leq \sigma$ holds. We call such $\mathbf{x}$'s *bad* samples since the values of h, and g are within the σ noise limit, and thus the output function $\widehat{p}$ can be close to both h and g. If all the samples are bad, the algorithm will not be able to distinguish between g and h, so $\widehat{p}$ is independent of the given samples. Since g and h are at least $1 > 2C\sigma$ far from each other, $\widehat{p}$ can well approximate at most one of them, which success with probability at most $1/2$ — the best the algorithm can do is to guess h, or g. We next show that all the samples are bad with high probability,

$$\Pr[|g(\mathbf{x}) - h(\mathbf{x})| \leq \sigma, \text{ for all samples}] \geq 1 - M \cdot \Pr_{\mathbf{x}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq \sigma\right] \geq 1 - Me^{-n\sigma^2/2} > 2/3.$$

The first inequality is by a union bound, the second is an application of Hoeffding's inequality. The last inequality then follows by the assumption that $M = e^{o(n\sigma^2)} < \frac{1}{3}e^{n\sigma^2/2}$. □

We note that both the Chebyshev and the uniform distribution are product distributions with mean 0. The constraint on the distribution comes from using Hoeffding's inequality. In general, Theorem A.1 holds for samples $\mathbf{x}$, where the coordinates x_i are independent, with $\mathbb{E}[x_i] = 0$ for all i.

B Deferred Proofs From Section 3

Proof of Lemma 3.5. Let us define three piece-wise constant functions with respect to the Chebyshev partition $r, \widehat{r}$, and $\tilde{r}$ as follows: For every $\boldsymbol{j} \in [m]^n$ and for every $\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}$,

$$\widehat{r}(\mathbf{x}) = \widehat{p}(\tilde{\mathbf{x}}_{\boldsymbol{j}}) \text{ , } \tilde{r}(\mathbf{x}) = \tilde{y}_{\boldsymbol{j}}, \text{ and } r(\mathbf{x}) = p(\mathbf{x}'_{\boldsymbol{j}})$$

where $\mathbf{x}'_{\boldsymbol{j}} \in \mathcal{C}_{\boldsymbol{j}}$ is chosen such that $(\mathbf{x}'_{\boldsymbol{j}}, \tilde{y}_{\boldsymbol{j}})$ is an inlier sample. The existence of such a point $\mathbf{x}'_{\boldsymbol{j}} \in \mathcal{C}_{\boldsymbol{j}}$ follows from the continuity of p and $\alpha < 0.5$, i.e. more than half of the samples in $\mathcal{C}_{\boldsymbol{j}}$ being inliers. Formally:

Claim B.1. *For every $\boldsymbol{j} \in [m]^n$, there exists a point $\mathbf{x}'_{\boldsymbol{j}} \in \mathcal{C}_{\boldsymbol{j}}$, such that $|p(\mathbf{x}'_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}| \leq \sigma$.*

Proof. Fix some $\boldsymbol{j} \in [m]^n$. Since, $\tilde{y}_{\boldsymbol{j}}$ is the median of all the y_i's, whose corresponding $\mathbf{x}_i \in \mathcal{C}_{\boldsymbol{j}}$, there must exist $\mu, \tau \in M$, such that $\tilde{y}_{\boldsymbol{j}} \in [y_\mu, y_\tau]$. Also, since more than half of the samples in $\mathcal{C}_{\boldsymbol{j}}$ are inliers, we can additionally force $(\mathbf{x}_\mu, y_\mu)$, and $(\mathbf{x}_\tau, y_\tau)$ to be inliers, i.e.

$$|p(\mathbf{x}_\mu) - y_\mu| \leq \sigma, \text{ and } |p(\mathbf{x}_\tau) - y_\tau| \leq \sigma.$$

So, we get $\tilde{y}_{\boldsymbol{j}} \in [p(\mathbf{x}_\mu) + \sigma, p(\mathbf{x}_\tau) - \sigma]$, and depending on whether $\tilde{y}_{\boldsymbol{j}}$ is closer to $p(\mathbf{x}_\mu) + \sigma$, or $p(\mathbf{x}_\tau) - \sigma$, we may choose $\mathbf{x}'_{\boldsymbol{j}}$ to be $\mathbf{x}_\mu$, or $\mathbf{x}_\tau$, respectively, ensuring $|p(\mathbf{x}'_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}| \leq \sigma$. □

Thus,

$$\|r - \tilde{r}\|_{\mathcal{C}_n,\infty} = \max_{\boldsymbol{j}\in[m]^n} |\tilde{y}_{\boldsymbol{j}} - p(\mathbf{x}'_{\boldsymbol{j}})| \leq \sigma. \tag{22}$$

Since $p \in \mathcal{P}_{\boldsymbol{d}}$ and $\widehat{p}$ is the minimizer of (1), we have,

$$\|\widehat{r} - \tilde{r}\|_{\mathcal{C}_n,\infty} = \max_{\boldsymbol{j}\in[m]^n} |\widehat{p}(\tilde{\mathbf{x}}_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}| \leq \max_{\boldsymbol{j}\in[m]^n} |p(\tilde{\mathbf{x}}_{\boldsymbol{j}}) - \tilde{y}_{\boldsymbol{j}}| \leq \|p - \tilde{r}\|_{\mathcal{C}_n,\infty} \tag{23}$$

Further, by the triangle inequality, we have

$$\|p - \widehat{p}\|_{\mathcal{C}_n,\infty} \leq \|p - \tilde{r}\|_{\mathcal{C}_n,\infty} + \|\tilde{r} - \widehat{r}\|_{\mathcal{C}_n,\infty} + \|\widehat{r} - \widehat{p}\|_{\mathcal{C}_n,\infty}$$

$$
\begin{aligned}
&\leq 2\|p-\tilde{r}\|_{\mathcal{C}_n,\infty} + \varepsilon\|\widehat{p}\|_{\mathcal{C}_n,\infty} && \text{(By (23) and Theorem 1.8 for } \widehat{p}\text{)}\\
&\leq 2(\|p-r\|_{\mathcal{C}_n,\infty} + \|r-\tilde{r}\|_{\mathcal{C}_n,\infty}) + \varepsilon(\|p-\widehat{p}\|_{\mathcal{C}_n,\infty} + \|p\|_{\mathcal{C}_n,\infty})\\
&\leq 3\varepsilon\|p\|_{\mathcal{C}_n,\infty} + 2\sigma + \varepsilon\|p-\widehat{p}\|_{\mathcal{C}_n,\infty} && \text{(By Theorem 1.8 for } p \text{ and (22))}
\end{aligned}
$$

Rearranging, and using the fact that $\frac{1}{1-\varepsilon} \leq 1 + 2\varepsilon \leq 2$ for $\varepsilon \leq 1/2$, we conclude:

$$\|p-\widehat{p}\|_{\mathcal{C}_n,\infty} \leq (2+4\varepsilon)\sigma + 6\varepsilon\|p\|_{\mathcal{C}_n,\infty}.$$

Rescaling ε to $\varepsilon/6$ gives the desired bound. □

C Deferred Proofs From Section 6

In order to prove Theorem 6.3, we need a variation (which is, in fact, a corollary) of Theorem 1.8 for ℓ_1:

Corollary C.1 (ℓ_1 approximation by piece-wise constant functions)**.** *Let $p : \mathcal{C} \to \mathbb{R}$ be a polynomial of individual degree at most d, and $r\colon \mathcal{C} \to \mathbb{R}$ a piece-wise constant function with respect to the (m,n)-Chebyshev partition, such that for all $\boldsymbol{j} \in [m]^n$ there exists $\mathbf{x}^{\boldsymbol{j}} \in \mathcal{C}_{\boldsymbol{j}}$, for which: $r(\mathbf{x}) = p(\mathbf{x}^{\boldsymbol{j}})$, for all $\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}$. Then, for some absolute constant $c > 1$,*

$$\|p-r\|_{\mathcal{C}_n,1} \leq \frac{(cd)^{2n+1}}{m}\|p\|_{\mathcal{C}_n,1}.$$

Before proving it, we note a useful result, and an observation:

Lemma C.2. *[Hölder's Inequality] Let $\alpha, \beta, \gamma \in \mathbb{R}_{\geq 1}$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = \frac{1}{\gamma}$. For all functions f and g with finite $\|f\|_{S,\alpha}$, and $\|g\|_{S,\beta}$, we have: $\|fg\|_{S,\gamma} \leq \|f\|_{S,\alpha}\|g\|_{S,\beta}$.*

Observation C.3. *[Equivalence of norms] Let $1 \leq \gamma < \alpha$, and $S \subseteq \mathcal{C}$. Define $\frac{1}{\beta} \triangleq \frac{1}{\gamma} - \frac{1}{\alpha}$, and $g(\mathbf{x}) \triangleq 1$, for all $\mathbf{x} \in \mathcal{C}$. Then, for any $f : \mathcal{C} \to \mathbb{R}$ with finite $\|f\|_{S,\alpha}$, by Lemma C.2, we have*

$$\|f\|_{S,\gamma} \leq \|f\|_{S,\alpha}\|\mathbf{1}\|_{S,\beta} = V_n^{\frac{1}{\beta}}(S)\|f\|_{S,\alpha} \leq 2^{\frac{n}{\beta}}\|f\|_{S,\alpha} = 2^{\frac{n}{\gamma}-\frac{n}{\alpha}}\|f\|_{S,\alpha}.$$

In particular, in the limit of $\alpha \to \infty$, we have

$$\|f\|_{S,\gamma} \leq V_n^{\frac{1}{\gamma}}(S)\|f\|_{S,\infty} \leq 2^{\frac{n}{\gamma}}\|f\|_{S,\infty}. \tag{24}$$

Proof of Corollary C.1. Observe,

$$\|p-r\|_{\mathcal{C}_n,1} \leq 2^n\|p-r\|_{\mathcal{C}_n,\infty} \leq O\left(\frac{dn}{m}\right)2^n\|p\|_{\mathcal{C}_n,\infty} \leq O\left(\frac{dn}{m}\right)(2\sqrt{2}d)^{2n}\|p\|_{\mathcal{C}_n,1}.$$

The first inequality is from Observation C.3(24) with $\gamma = 1$, the second is by Theorem 1.8, and the last by Theorem 1.9. □

We next generalize [KKP17, Lemma 3.1]. We show that on a fine enough Chebyshev grid, an empirical estimate of the ℓ_1 norm of p is close to the actual value, thus implying it can be used as a proxy for the actual:

Theorem C.4 (Empirical ℓ_1 estimate suffices)**.** *Let $p : \mathcal{C} \to \mathbb{R}$ be a polynomial of individual degree at most d, and $m \geq (cd)^{2n+1}/\varepsilon$, for large enough constant $c > 1$. Given a set of M samples $(\mathbf{x}_i, y_i)$, such that for every $\boldsymbol{j} \in [m]^n$, the set of samples in the cell $\mathcal{C}_{\boldsymbol{j}}$, denoted by $S_{\boldsymbol{j}} \triangleq \{\alpha \in [M] : \mathbf{x}_\alpha \in \mathcal{C}_{\boldsymbol{j}}\}$ is not empty. Define the weighted average of p with respect to the the (m, n)-Chebyshev partition:*

$$\Phi(p) \triangleq \sum_{\boldsymbol{j} \in [m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\alpha \in S_{\boldsymbol{j}}} |p(\mathbf{x}_\alpha)|.$$

Then, $\Phi(p) \in (1 \pm \varepsilon)\|p\|_{\mathcal{C}_n,1}$.

Proof. Fix some $\boldsymbol{j} \in [m]^n$. Denote the average value of samples in the cell $\mathcal{C}_{\boldsymbol{j}}$ by $r_{\boldsymbol{j}} \triangleq \sum_{\alpha \in S_{\boldsymbol{j}}} \frac{|p(\mathbf{x}_\alpha)|}{|S_{\boldsymbol{j}}|}$. Define

$$a_{\boldsymbol{j}} \triangleq \min_{\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}} \{|p(\mathbf{x})|\}, \qquad \text{and} \qquad b_{\boldsymbol{j}} \triangleq \max_{\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}} \{|p(\mathbf{x})|\}.$$

Observe $r_{\boldsymbol{j}} \in [a_{\boldsymbol{j}}, b_{\boldsymbol{j}}]$. By continuity of $|p|$, there exists $\mathbf{x}^{(\boldsymbol{j})} \in \mathcal{C}_{\boldsymbol{j}}$, such that $|p(\mathbf{x}^{(\boldsymbol{j})})| = r_{\boldsymbol{j}}$.

Let $r \colon \mathcal{C} \to \mathbb{R}$ be a piece-wise constant function with respect to the (m, n)-Chebyshev partition, which is defined as: $r(\mathbf{x}) = p(\mathbf{x}^{\boldsymbol{j}})$, for all $\mathbf{x} \in \mathcal{C}_{\boldsymbol{j}}$. Observe,

$$\Phi(p) = \sum_{\boldsymbol{j} \in [m]^n} V_n(\mathcal{C}_{\boldsymbol{j}}) r_{\boldsymbol{j}} = \int_{\mathcal{C}} |r(\mathbf{x})| d\mathbf{x} = \|r\|_{\mathcal{C}_n,1}.$$

By triangle inequality, we have

$$\|p\|_{\mathcal{C}_n,1} - \|r - p\|_{\mathcal{C}_n,1} \leq \Phi(p) = \|r\|_{\mathcal{C}_n,1} \leq \|r - p\|_{\mathcal{C}_n,1} + \|p\|_{\mathcal{C}_n,1}$$

By our choice of m, and Corollary C.1, we have $\|p - r\|_{\mathcal{C}_n,1} \leq \varepsilon\|p\|_{\mathcal{C}_n,1}$. So, we conclude

$$(1 - \varepsilon)\|p\|_{\mathcal{C}_n,1} \leq \Phi(p) \leq (1 + \varepsilon)\|p\|_{\mathcal{C}_n,1}.$$ □

Next, we define the *closeness* of a set of samples to p, in ℓ_1, and relate it to our notion of *goodness*. We show our ℓ_1 minimizer outputs a close approximation, if the samples are good (and, hence close to p in ℓ_1).

Definition C.5 ((α, γ)-close in ℓ_1)**.** *Let $p \colon \mathcal{C} \to \mathbb{R}$ be a polynomial of individual degree at most d. For a set of M samples $S \triangleq \{(\mathbf{x}_i, y_i)\}$, consider the (m, n)-Chebyshev partition. Let $S_{\boldsymbol{j}} \triangleq \{\beta \in [M] : \mathbf{x}_\beta \in \mathcal{C}_{\boldsymbol{j}}\}$ be the set of samples that are in the cell $\mathcal{C}_{\boldsymbol{j}}$. Let $\alpha < 1/2$, and $\gamma > 0$. For every $\boldsymbol{j} \in [m]^n$, let*

$$e_{\boldsymbol{j}} \triangleq \min_{\substack{S' \subset S_{\boldsymbol{j}}, \\ |S'| \leq \lceil (1-\alpha)|S_{\boldsymbol{j}}| \rceil}} \max_{\beta \in S'} |p(\mathbf{x}_\beta) - y_\beta|.$$

We say that S is (α, γ)-close to p in ℓ_1 with respect to the partition, if $|S_{\boldsymbol{j}}| \geq 1/\alpha$ for all $\boldsymbol{j} \in [m]^n$ and, in addition,

$$\sum_{\boldsymbol{j} \in [m]^n} V_n(\mathcal{C}_{\boldsymbol{j}}) e_{\boldsymbol{j}} \leq \gamma. \tag{25}$$

For every $\boldsymbol{j} \in [m]^n$, let $S'_{\boldsymbol{j}} \subseteq S_{\boldsymbol{j}}$ be the set of inliers in the cell $\mathcal{C}_{\boldsymbol{j}}$. If a set of samples S is α-good, then the fraction of outliers in $S_{\boldsymbol{j}}$ is less than α, and hence $|S'_{\boldsymbol{j}}| > (1 - \alpha)|S_{\boldsymbol{j}}|$. Note that for every $\beta \in S'_{\boldsymbol{j}}$, $|p(\mathbf{x}_\beta) - y_\beta| \leq \sigma$, making $e_{\boldsymbol{j}} \leq \sigma$. Hence $\sum_{\boldsymbol{j} \in [m]^n} V_n(\mathcal{C}_{\boldsymbol{j}}) e_{\boldsymbol{j}} \leq \sigma V_n(\mathcal{C}) = 2^n \sigma$. We conclude:

Observation C.6 (α-good $\implies$ $(\alpha, 2^n\sigma)$-close)**.** *If a set S of samples is α-good for the (m, n)-Chebyshev partition, then S is also $(\alpha, 2^n\sigma)$-close to p in ℓ_1.*

Finally, we bound the ℓ_1 regression error of the ℓ_1 minimizer $\widehat{p}_{\ell_1}$, with respect to p. The minimizer $\widehat{p}_{\ell_1}$ is assumed to be computed on a set of samples S, that is (α,γ)-close to p in ℓ_1:

Lemma C.7 (ℓ_1 error bound for the ℓ_1 minimizer)**.** *Let $\alpha < 1/2, \varepsilon \le (1-2\alpha)/2$, and $m \ge (cd)^{2n+1}/\varepsilon$, for some large enough constant $c > 1$. Given a set of M samples $S \triangleq \{(\mathbf{x}_i, y_i)\}$, that is (α,γ)-close to an individual degree-d polynomial $p\colon \mathcal{C} \to \mathbb{R}$ in ℓ_1 with respect to the (m,n)-Chebyshev grid, if $\widehat{p}_{\ell_1}$ is the ℓ_1 minimizer from Definition 6.2, then*

$$\|p - \widehat{p}_{\ell_1}\|_{\mathcal{C}_n,1} \le \frac{4\gamma}{1-2\alpha}.$$

Proof. In each cell $\mathcal{C}_{\boldsymbol{j}}$, call the $\lfloor \alpha |S_{\boldsymbol{j}}| \rfloor$ points that maximize $|p(\mathbf{x}_\beta) - y_\beta|$ the *bad* points, and the rest are *good*, denoted by $B_{\boldsymbol{j}}$, and $G_{\boldsymbol{j}}$ respectively. Let the objective function be defined as

$$obj(f) \triangleq \sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\beta\in S_{\boldsymbol{j}}} |f(\mathbf{x}_\beta) - y_\beta|,$$

and with $\widehat{p}_{\ell_1}$ as its minimizer, we have

$$0 \ge obj(\widehat{p}_{\ell_1}) - obj(p) = \sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\beta\in S_{\boldsymbol{j}}} (|\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - y_\beta| - |p(\mathbf{x}_\beta) - y_\beta|) = (\star).$$

Now, using the triangle inequality, for the *good* samples, we bound

$$|\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - y_\beta| \ge |\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - p(\mathbf{x}_\beta)| - |p(\mathbf{x}_\beta) - y_\beta|.$$

Whereas for the *bad* samples, we bound

$$|\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - y_\beta| - |p(\mathbf{x}_\beta) - y_\beta| \ge -|p(\mathbf{x}_\beta) - \widehat{p}_{\ell_1}(\mathbf{x}_\beta)|.$$

Therefore,

$$\begin{aligned}(\star) &\ge \sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\beta\in G_{\boldsymbol{j}}} (|\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - p(\mathbf{x}_\beta)| - 2|p(\mathbf{x}_\beta) - y_\beta|) - \sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\beta\in B_{\boldsymbol{j}}} |\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - p(\mathbf{x}_\beta)| \\ &\ge \underbrace{\sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\beta\in G_{\boldsymbol{j}}} |\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - p(\mathbf{x}_\beta)|}_{(I)} - \underbrace{\sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} \sum_{\beta\in B_{\boldsymbol{j}}} |\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - p(\mathbf{x}_\beta)|}_{(II)} - \underbrace{2 \sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|S_{\boldsymbol{j}}|} |G_{\boldsymbol{j}}| e_{\boldsymbol{j}}}_{(III)},\end{aligned}$$

where last inequality follows, since $|p(\mathbf{x}_\beta) - y_\beta| \le e_{\boldsymbol{j}}$ for all *good* samples. Next, we bound each term separately. The third term is $(III) \le \gamma$ follows from (25) and the fact that $G_{\boldsymbol{j}} \subseteq S_{\boldsymbol{j}}$.

For the first and second terms, we use the fact that for every $\boldsymbol{j} \in [m]^n$, $|B_{\boldsymbol{j}}| \le \alpha |S_{\boldsymbol{j}}|$ and $|G_{\boldsymbol{j}}| \ge (1-\alpha)|S_{\boldsymbol{j}}|$, we note that since $|S_{\boldsymbol{j}}| \ge 1/\alpha$, both $B_{\boldsymbol{j}}$ and $G_{\boldsymbol{j}}$ are not empty, and we can apply Theorem C.4.

$$(I) \ge (1-\alpha) \sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|G_{\boldsymbol{j}}|} \sum_{\beta\in G_{\boldsymbol{j}}} |\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - p(\mathbf{x}_\beta)| \ge (1-\alpha)(1-\varepsilon)\|\widehat{p}_{\ell_1} - p\|_{\mathcal{C}_n,1},$$

and

$$(II) \le \alpha \sum_{\boldsymbol{j}\in[m]^n} \frac{V_n(\mathcal{C}_{\boldsymbol{j}})}{|B_{\boldsymbol{j}}|} \sum_{\beta\in B_{\boldsymbol{j}}} |\widehat{p}_{\ell_1}(\mathbf{x}_\beta) - p(\mathbf{x}_\beta)| \le \alpha(1+\varepsilon)\|\widehat{p}_{\ell_1} - p\|_{\mathcal{C}_n,1}.$$

Combining the bounds and rearranging we conclude:

$$\|p - \widehat{p}_{\ell_1}\|_{\mathcal{C}_n,1} \le \frac{2\gamma}{1-2\alpha-\varepsilon} \le \frac{4\gamma}{1-2\alpha}.$$

The last inequality follows from the assumption $\varepsilon \le (1-2\alpha)/2$. □

As an immediate corollary, we can bound the ℓ_∞ error of the ℓ_1 minimizer $\widehat{p}_{\ell_1}$ on a set S of samples that is α-good, thus proving the main theorem of this subsection.

Proof of Theorem 6.3. Since S is assumed to be α-good, by Observation C.6, it is $(\alpha, 2^n\sigma)$-close to p in ℓ_1. So, invoking Lemma C.7, we get $\|p - \widehat{p}_{\ell_1}\|_{\mathcal{C}_n,1} \leq 2^n O_\alpha(\sigma)$. Further, since both p, and $\widehat{p}_{\ell_1}$ are polynomials of individual degree at most d, so is their difference. Hence, invoking Theorem 1.9, we get $\|p - \widehat{p}_{\ell_1}\|_{\mathcal{C}_n,\infty} \leq (4d^2)^n \|p - \widehat{p}_{\ell_1}\|_{\mathcal{C}_n,1} \leq (8d^2)^n O_\alpha(\sigma)$. □

D Weaker ℓ_∞ to ℓ_1 norms relation

This section is devoted to proving Theorem D.3, a weaker version of Theorem 1.9 for *total* degree-d polynomials, the proof of which relies on a result by Wilhelmsen [Wil74] and avoids the inductive argument of Theorem 1.9. Using this weaker bound instead would lead to sample complexity of $O(m^n \log m^n) \geq ((c_0 d)^{2n+2} n^{n/2})^n \log d$ instead of $(cd)^{n(2n+1)} \log d$, for some absolute constants $c, c_0 > 0$.

Let T be a compact, convex subset of $\mathbb{R}^n$, with boundary ∂T, and interior $T^0 \neq \emptyset$. Fix a $\boldsymbol{t}_0 \in \partial T$ and a unit vector $\boldsymbol{u} \in \mathbb{R}^n$. Consider the hyperplane with normal $\boldsymbol{u}$

$$\mathcal{H}_{\boldsymbol{u}} \triangleq \{\boldsymbol{t} \in \mathbb{R}^n : \langle \boldsymbol{t} - \boldsymbol{t}_0, \boldsymbol{u} \rangle = 0\}.$$

$\mathcal{H}_{\boldsymbol{u}}$ is a support hyperplane of T at t_0 if is not containing any $\boldsymbol{t} \in T^0$. In that case, $\boldsymbol{u}$ is called an outer normal to T at t_0. For any direction $\boldsymbol{u}$, there exist precisely two support hyperplanes of T, with outer normals $\boldsymbol{u}$, and $-\boldsymbol{u}$. They are separated by a distance $\lambda_{\boldsymbol{u}} > 0$. The width of T is defined ([Wil74, Definition 2.1]) to be $\omega_T \triangleq \min_{\|\boldsymbol{u}\|_2=1} \lambda_{\boldsymbol{u}}$. We will use the following result by Wilhelmsen:

Theorem D.1. *[Wil74, Theorem 3.1] Let T be a compact, convex subset of $\mathbb{R}^n$. For any* total *degree-d polynomial $p : T \to \mathbb{R}$, we have*[12]

$$\|\boldsymbol{\nabla} p\|_{T,\infty} \leq \frac{4d^2}{\omega_T} \|p\|_{T,\infty}.$$

Observe, the width of $\mathcal{C} = [-1, 1]^n$ is $\omega_{\mathcal{C}} = 2$. So, we have

Observation D.2. *[Wilhelmsen on the cube $\mathcal{C}_n$]*

$$\|\boldsymbol{\nabla} p\|_{\mathcal{C}_n,\infty} \leq 2d^2 \|p\|_{\mathcal{C}_n,\infty}$$

Using this observation, we prove a weaker version of Theorem 1.9:

Theorem D.3. *[Weaker ℓ_∞-ℓ_1 norms relation] Let $p : \mathcal{C} \to \mathbb{R}$ be a polynomial of* total *degree-d. Then,*

$$\|p\|_{\mathcal{C}_n,\infty} \leq 2\sqrt{n\pi} \left(4d^2 \sqrt{\frac{2n}{e\pi}}\right)^n \|p\|_{\mathcal{C}_n,1}.$$

Proof. Let $\mathbf{x}^* \in \mathcal{C} : p(\mathbf{x}^*) = \|p\|_{\mathcal{C}_n,\infty}$. Define a set of points close enough to $\mathbf{x}^*$, i.e. $S(\mathbf{x}^*) \triangleq \{\boldsymbol{y} \in \mathcal{C} : \|\mathbf{x}^* - \boldsymbol{y}\|_2 \leq 1/4d^2\} \subseteq \mathcal{C}$. Then, we lower bound p on all points in $S(\mathbf{x}^*)$. Formally:

Lemma D.4. *For every $\boldsymbol{y} \in S(\mathbf{x}^*)$, $|p(\boldsymbol{y})| \geq |p(\mathbf{x}^*)|/2$.*

[12] Here, for any $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|$ denotes $\|\boldsymbol{v}\|_2$, the ℓ_2 norm of $\boldsymbol{v}$, and $\|\boldsymbol{v}\|_{T,\infty}$ denotes $\max_{\boldsymbol{v} \in T}\{\|\boldsymbol{v}\|_2\}$.

Proof. Fix a $\boldsymbol{y} \in S(\mathbf{x}^*)$. Let $\mathrm{L}_{\mathbf{x}^*,\boldsymbol{y}}$ be the line segment connecting $\mathbf{x}^*$ and $\boldsymbol{y}$. Then, by Theorem 5.2, there exists a $\boldsymbol{z} \in \mathrm{L}_{\mathbf{x}^*,\boldsymbol{y}-\mathbf{x}^*}$, such that $p(\mathbf{x}^*) - p(\boldsymbol{y}) = \langle \boldsymbol{\nabla} p(\boldsymbol{z}), \mathbf{x}^* - \boldsymbol{y} \rangle$.

$$\begin{aligned}
\implies |p(\mathbf{x}^*) - p(\boldsymbol{y})| &= |\langle \boldsymbol{\nabla} p(\boldsymbol{z}), \mathbf{x}^* - \boldsymbol{y} \rangle| \leq \|\boldsymbol{\nabla} p(\boldsymbol{z})\|_2 \cdot \|\mathbf{x}^* - \boldsymbol{y}\|_2 && \text{(By Cauchy-Schwarz)} \\
&\leq \|\boldsymbol{\nabla} p\|_{\mathcal{C}_{n,\infty}} \cdot \underbrace{\|\mathbf{x}^* - \boldsymbol{y}\|_2}_{\leq 1/4d^2, \because \boldsymbol{y} \in S(\mathbf{x}^*)} \\
&\leq 2d^2 \|p\|_{\mathcal{C}_{n,\infty}} / (4d^2) && \text{(By Observation D.2)} \\
&= \frac{\|p\|_{\mathcal{C}_{n,\infty}}}{2} = \frac{|p(\mathbf{x}^*)|}{2}.
\end{aligned}$$

$$\implies |p(\boldsymbol{y})| \geq |p(\mathbf{x}^*)| - |p(\mathbf{x}^*) - p(\boldsymbol{y})| \geq |p(\mathbf{x}^*)|/2$$ □

Now observe,

$$\begin{aligned}
\|p\|_{\mathcal{C}_{n,1}} = \int_{\mathcal{C}} |p(\mathbf{y})| d\mathbf{y} &\geq \int_{S(\mathbf{x}^*) \cap \mathcal{C}} |p(\boldsymbol{y})| d\boldsymbol{y} \geq \int_{S(\mathbf{x}^*) \cap \mathcal{C}} \frac{|p(\mathbf{x}^*)|}{2} d\boldsymbol{y} && \text{(By Lemma D.4)} \\
&\geq \frac{1}{2^n} \int_{S(\mathbf{x}^*)} \frac{|p(\mathbf{x}^*)|}{2} d\mathbf{x} && (\, S(\mathbf{x}^*) \cap \mathcal{C} \text{ covers at least one orthant of } S(\mathbf{x}^*)) \\
&= \frac{|p(\mathbf{x}^*)|}{2^{n+1}} V_n(S(\mathbf{x}^*)) = \frac{|p(\mathbf{x}^*)|}{2^{n+1}} \cdot \frac{\pi^{n/2}}{\Gamma(n/2+1)(4d^2)^n} && (S(\mathbf{x}^*) \text{ is an } n\text{-ball of radius } 1/4d^2) \\
&\approx \frac{\|p\|_{\mathcal{C}_{n,\infty}}}{2^{n+1}\sqrt{n\pi}} \left(\frac{1}{4d^2} \sqrt{\frac{2e\pi}{n}} \right)^n. && \text{(By Stirling's approximation of the Gamma function)}
\end{aligned}$$

$$\implies \|p\|_{\mathcal{C}_{n,\infty}} \leq 2\sqrt{n\pi} \left(4d^2 \sqrt{\frac{2n}{e\pi}} \right)^n \|p\|_{\mathcal{C}_{n,1}}.$$ □

If we use Theorem D.3, instead of Theorem 1.9, we get a weaker form of Corollary C.1:

Corollary D.5. *Let $p : \mathcal{C} \to \mathbb{R}$ be a polynomial of* total *degree at most d, and $r : \mathcal{C} \to \mathbb{R}$ a piece-wise constant function with respect to the (m,n)-Chebyshev partition, such that for all $\boldsymbol{j} \in [m]^n$ there exists $\mathbf{x}^{\boldsymbol{j}} \in \mathcal{C}_{\boldsymbol{j}}$, for which $r(\mathbf{x}) = p(\mathbf{x}^{\boldsymbol{j}})$. Then*

$$\|p - r\|_{\mathcal{C}_{n,1}} \leq O\left(\frac{d^2 n \sqrt{n}}{m} (16 d^2 \sqrt{n})^n \right) \|p\|_{\mathcal{C}_{n,1}}.$$

Using this, we get that, for $m \geq \frac{cd^2 n\sqrt{n}}{\varepsilon}(16d^2\sqrt{n})^n$, for some absolute constant $c > 0$, the ℓ_∞ error of the ℓ_1 minimizer $\widehat{p}_{\ell_1}$, with respect to the (m,n)-Chebyshev grid, can be bounded by $poly(d^n)\sigma$, via a worse form of Theorem 6.3:

Corollary D.6. *Let $\alpha < 0.5$ be constant , $\varepsilon \leq (1-2\alpha)/2$, and $m \geq \frac{c_0^n d^{2n+2} n^{n/2}}{\varepsilon} \geq \frac{cd^2 n\sqrt{n}}{\varepsilon}(16d^2\sqrt{n})^n$, for some constants $c > 0, c_0 > 0$. Given a set S of that is α-good with respect to the (m,n)-Chebyshev partition, with $\widehat{p}_{\ell_1}$ as in Definition 6.2, we have*[13]

$$\|p - \widehat{p}_{\ell_1}\|_{\mathcal{C}_{n,\infty}} \leq O((8d^2)^n \sigma).$$

Since we use ℓ_1 minimizer, only in the case of arbitrarily large $\|p\|_{\mathcal{C}_{n,\infty}}$ (Section 6), this worsens only the sample complexity of Algorithm 3.

[13] Here p, and $\widehat{p}_{\ell_1}$ are polynomials of *total* degree at most d.