Robust polynomial regression up to the information theoretic limit

Daniel Kane  Sushrut Karmalkar  Eric Price

August 11, 2017

Abstract

We consider the problem of robust polynomial regression, where one receives samples \((x_i, y_i)\) that are usually within \(\sigma\) of a polynomial \(y = p(x)\), but have a \(\rho\) chance of being arbitrary adversarial outliers. Previously, it was known how to efficiently estimate \(p\) only when \(\rho < \frac{1}{\log d}\). We give an algorithm that works for the entire feasible range of \(\rho < \frac{1}{2}\), while simultaneously improving other parameters of the problem. We complement our algorithm, which gives a factor 2 approximation, with impossibility results that show, for example, that a 1.09 approximation is impossible even with infinitely many samples.

1 Introduction

Polynomial regression is the problem of finding a polynomial that passes near a collection of input points. The problem has been studied for 200 years with diverse applications [Ger74], including computer graphics [Pra87], machine learning [KKMS08], and statistics [Mac78]. As with linear regression, the classic solution to polynomial regression is least squares, but this is not robust to outliers: a single outlier point can perturb the least squares solution by an arbitrary amount. Hence finding a "robust" method of polynomial regression is an important question.

A version of this problem was formalized by Arora and Khot [AK03]. We want to learn a degree \(d\) polynomial \(p: [-1, 1] \to \mathbb{R}\), and we observe \((x_i, y_i)\) where \(x_i\) is drawn from some measure \(\chi\) (say, uniform) and \(y_i = p(x_i) + w_i\) for some noise \(w_i\). Each sample is an "outlier" independently with probability at most \(\rho\); if the sample is not an outlier, then \(|w_i| \leq \sigma\) for some \(\sigma\). Other than the random choice of outlier locations, the noise is adversarial. One would like to recover \(\hat{p}\) with \(\|p - \hat{p}\|_{\infty} \leq C\sigma\) using as few samples as possible, with high probability.

In other contexts an \(\ell_\infty\) requirement on the input noise would be a significant restriction, but outlier tolerance makes it much less so. For example, independent bounded-variance noise fits in this framework: by Chebyshev’s inequality, the framework applies with outlier chance \(\rho = \frac{1}{100}\) and \(\sigma = 10\mathbb{E}[w_i^2]^{1/2}\), which gives the ideal result up to constant factors. At the same time, the \(\ell_\infty\) requirement on the input allows for the strong \(\ell_\infty\) bound on the output.

Arora and Khot [AK03] showed that \(\rho < 1/2\) is information-theoretically necessary for non-trivial recovery, and showed how to solve the problem using a linear program when \(\rho = 0\). For \(\rho > 0\), the RANSAC [FB81] technique of fitting to a subsample of the data works in polynomial time when \(\rho \lesssim \frac{\log d}{d}\). Finding an efficient algorithm for \(\rho \gg \frac{\log d}{d}\) remained open until recently.

Last year, Guruswami and Zuckerman [GZ16] showed how to solve the problem in polynomial time for \(\rho\) larger than \(\frac{\log d}{d}\). Unfortunately, their result falls short of the ideal in several ways: it needs a low outlier rate \((\rho < \frac{1}{\log d})\), a bounded signal-to-noise ratio \((\frac{|p|_{\infty}}{\sigma}) < d^{O(1)})\), and has a super-constant approximation factor \((\|p - \hat{p}\|_{\infty} \lesssim \sigma \left(1 + \frac{|p|_{\infty}}{\sigma}\right)^{0.01})\). It uses \(O(d^2 \log^c d)\) samples from the uniform measure, or \(O(d \log^c d)\) samples from the Chebyshev measure \(\frac{1}{\sqrt{1-x^2}}\).
These deficiencies mean that the algorithm doesn’t work when all the noise comes from outliers ($\sigma = 0$); the low outlier rate required also leads to a superconstant factor loss when reducing from other noise models.

In this work, we give a simple algorithm that avoids all these deficiencies: it works for all $\rho < 1/2$; it has no restrictions on $\sigma$; and it gets a constant approximation factor $C = 2 + \epsilon$. Additionally, it only uses $O(d^2)$ samples from the uniform measure, without any log factors, and $O(d \log d)$ samples from the Chebyshev measure. We also give lower bounds for the approximation factor $C$, indicating that one cannot achieve a $1 + \epsilon$ approximation by showing that $C > 1.09$. Our lower bounds are not the best possible, and it is possible that the true lower bounds are much better.

### 1.1 Algorithmic results

The problem formulation has randomly placed outliers, which is needed to ensure that we can estimate the polynomial locally around any point $x$. We use the following definition to encapsulate this requirement, after which the noise can be fully adversarial:

**Definition 1.1.** The size $m$ Chebyshev partition of $[-1, 1]$ is the set of intervals $I_j = [\cos \frac{\pi j}{m}, \cos \frac{\pi (j-1)}{m}]$ for $j \in [m]$.

We say that a set $S$ of samples $(x_i, y_i)$ is “$\alpha$-good” for the partition if, in every interval $I_j$, strictly less than an $\alpha$ fraction of the points $x_i \in I_j$ are outliers.

The size $m$ Chebyshev partition is the set of intervals between consecutive extrema of the Chebyshev polynomial of degree $m$. Standard approximation theory recommends sampling functions at the roots of the Chebyshev polynomial, so intuitively a good set of samples will allow fairly accurate estimation near each of these “ideal” sampling points. All our algorithms will work for any set of $\alpha$-good samples, for $\alpha < \frac{1}{2}$ and sufficiently large $m$. The sample complexities are then what is required for $S$ to be good with high probability.

We first describe two simple algorithms that do not quite achieve the goal. We then describe how to black-box combine their results to get the full result.

**L1 regression.** Our first result is that $\ell_1$ regression almost solves the problem: it satisfies all the requirements, except that the resulting error $\|\tilde{p} - p\|$ is bounded in $\ell_1$ not $\ell_\infty$:

**Lemma 1.2.** Suppose the set $S$ of samples is $\alpha$-good for the size $m = O(d)$ Chebyshev partition, for constant $\alpha < \frac{1}{2}$. Then the result $\tilde{p}$ of $\ell_1$ regression

$$\arg \min_{\text{degree-}d \text{ polynomial } \tilde{p}} \sum_{i=1}^{n} |I_j| \text{mean}_{x_i \in I_j} |y_i - \tilde{p}(x_i)|$$

satisfies $\|\tilde{p} - p\|_1 \leq O_\alpha(1) \cdot \sigma$.

This has a nice parallel to the $d = 1$ case, where $\ell_1$ regression is the canonical robust estimator.

As shown by the Chebyshev polynomial in Figure 1, $\ell_1$ regression might not give a good solution to the $\ell_\infty$ problem. However, converting to an $\ell_\infty$ bound loses at most an $O(d^2)$ factor. This means this lemma is already useful for an $\ell_\infty$ bound: one of the requirements of [GZ16] was that $\|p\|_\infty \leq \sigma d^{O(1)}$. We can avoid this by first computing the $\ell_1$ regression estimate $\tilde{p}^{(\ell_1)}$ and then applying [GZ16] to the residual $p - \tilde{p}^{(\ell_1)}$, which always satisfies the condition.
Figure 1: The blue observations are within $\sigma = 1$ of the red zero polynomial at each point, but are closer in $\ell_1$ to the green Chebyshev polynomial. This demonstrates that $\ell_1$ regression does not solve the $\ell_\infty$ problem even for $\rho = 0$.

**Median-based recovery.** Our second result is for a median-based approach to the problem: take the median $\tilde{y}_j$ of the $y$ values within each interval $I_j$. Because the median is robust, this will lie within the range of inlier $y$ values for that interval. We don’t know which $x \in I_j$ it corresponds to, but this turns out not to matter too much: assigning each $\tilde{y}_j$ to an arbitrary $\tilde{x}_j \in I_j$ and applying non-robust $\ell_\infty$ regression gives a useful result.

**Lemma 1.3.** Let $\epsilon, \alpha < \frac{1}{2}$, and suppose the set $S$ of samples is $\alpha$-good for the size $m = O(d/\epsilon)$ Chebyshev partition. Let $\tilde{x}_j \in I_j$ be chosen arbitrarily, and let $\tilde{y}_j = \text{median}_{x_i \in I_j} y_i$. Then the degree $d$ polynomial $\hat{p}$ minimizing

$$\max_{j \in [m]} |\hat{p}(\tilde{x}_j) - \tilde{y}_j|$$

satisfies $\|\hat{p} - p\|_\infty \leq (2 + \epsilon)\sigma + \epsilon\|p\|_\infty$.

Without the additive $\epsilon\|p\|_\infty$ term, which comes from not knowing the $x \in I_j$ for $\tilde{y}_j$, this would be exactly what we want. If $\sigma \ll \|p\|_\infty$, however, this is not so good—but it still makes progress.

**Iterating the algorithms.** While neither of the above results is sufficient on its own, simply applying them iteratively on the residual left by previous rounds gives our full theorem. The result of $\ell_1$ regression is a $\tilde{p}^{(0)}$ with $\|p - \tilde{p}^{(0)}\|_\infty = O(d^2\sigma)$. Applying the median recovery algorithm to the residual points $(x_i, y_i - \tilde{p}^{(0)}(x_i))$, we will get $\tilde{p}'$ so that $\tilde{p}^{(1)} = \tilde{p}^{(0)} + \tilde{p}'$ has

$$\|\tilde{p}^{(1)} - p\|_\infty \leq (2 + \epsilon)\sigma + \epsilon \cdot O(d^2\sigma)$$

If we continue applying the median recovery algorithm to the remaining residual, the $\ell_\infty$ norm of the residual will continue to decrease exponentially. After $r = O(\log_{1/\epsilon} d)$ rounds we will reach

$$\|\tilde{p}^{(r)} - p\|_\infty \leq (2 + 4\epsilon)\sigma$$

as desired. Since each method can be computed efficiently with a linear program, this gives our main theorem:

\[1\] We remark that one could skip the initial $\ell_1$ minimization step, but the number of rounds would then depend on $\log(\frac{|p|_\infty}{\sigma})$, which could be unbounded. Note that this only affects running time and not the sample complexity.
Theorem 1.4. Let $\epsilon, \alpha < \frac{1}{2}$, and suppose the set $S$ of samples is $\alpha$-good for the size $m = O(d/\epsilon)$ Chebyshev partition. The result $\hat{p}$ of Algorithm 2 is a degree $d$ polynomial satisfying $\|\hat{p} - p\|_\infty \leq (2 + \epsilon)\sigma$. Its running time is that of solving $O(\log_2 d)$ linear programs, each with $O(d)$ variables and $O(|S|)$ constraints.

We now apply this to the robust polynomial regression problem, where we receive points $(x_i, y_i)$ such that each $x_i$ is drawn from some distribution, and with probability $\rho$ it is an outlier. An adversary then picks the $y_i$ such that, for each non-outlier $i$, $|y_i - p(x_i)| \leq \sigma$. We observe the following:

Corollary 1.5. Consider the robust polynomial regression problem for constant outlier chance $\rho < 1/2$, with points $x_i$ drawn from the Chebyshev distribution $D_c(x) \sim \frac{1}{\sqrt{1-x^2}}$. Then $O(\frac{d^2}{\epsilon^2} \log \frac{d}{\epsilon})$ samples suffice to recover with probability $1 - \delta$ a degree $d$ polynomial $\hat{p}$ with

$$\max_{x \in [-1, 1]} |p(x) - \hat{p}(x)| \leq (2 + \epsilon)\sigma.$$  

If $x_i$ is drawn from the uniform distribution instead, then $O(\frac{d^2}{\epsilon^2} \log \frac{d}{\epsilon})$ samples suffice for the same result. In both cases, the recovery time is polynomial in the sample size.

1.2 Impossibility results

We show limitations on improving any of the three parameters used in Corollary 1.5: the sample complexity, the approximation factor, and the requirement that $\rho < 1/2$.

Sample Complexity. Our result uses $O(d^2)$ samples from the uniform distribution and $O(d \log d)$ samples from the Chebyshev distribution. We show in Section 5.1 that both results are tight. In particular, we show that it is impossible to get any constant approximation with $o(d^2)$ samples from the uniform distribution or $o(d \log d)$ samples from the Chebyshev distribution.

Approximation factor. Our approximation factor is $2 + \epsilon$. Even with infinitely many samples and no outliers, can one do better than a $2$-approximation for $\ell_\infty$ regression? For comparison, in $\ell_2$ regression a $1 + \epsilon$ approximation is possible. For these lower bounds, it is convenient to consider the special case where the values $y_i$ are $y(x_i)$ for a function $y$ with $\|p - y\|_\infty \leq \sigma$. We show two lower bounds related to this question.

First, we show that $\ell_\infty$ projection—that is, the algorithm that minimizes $\|y - \hat{p}\|_\infty$ over degree-$d$ polynomials $\hat{p}$—can have error arbitrarily close to $2\sigma$. Since our algorithm is an outlier-tolerant version of $\ell_\infty$ projection, we should not expect to perform better.

Second, we show that no proper learning algorithm can achieve a $1 + \epsilon$ guarantee. In particular, we show for $d = 2$ that any algorithm with more than $2/3$ success probability must have $C > 1.09$; this is illustrated in Figure 2. For general $d$, we show that $C > 1 + \Omega(1/d^3)$.

List decoding. The $\rho < 1/2$ requirement is obviously required for uniquely decoding $p$, since for $\rho \geq 1/2$ the observations could come from a mixture of two completely different polynomials. But one could hope for a list decoding version of the result, outputting a small set of polynomials such that one is close to the true output. Unfortunately, [AK03] showed that for a $C$-approximate algorithm, such a set would require size at least $e^{\Omega(\sqrt{d}/C)}$ event for $\rho \geq 1/2$. We improve this lower bound to $e^{\Omega(d/C)}$. At least for constant $C$ and $\rho$, our result is tight: Theorem 1.4 implies that if no samples were outliers, then $m = O(d)$ samples would suffice. Hence picking $m$ samples will result in
Figure 2: The lower bound for 1.09-approximations via any recovery algorithm. We give three quadratics that all lie within $2\sigma$ of each other for all $x$, and hence the observed $y(x)$ can be identical regardless of which quadratic is $p$. The region containing points within $1.09\sigma$ of all three options just barely doesn’t contain a quadratic itself.

Algorithm 2 outputting a polynomial consistent with the data with $(1 - \rho)^m = e^{-O(d)}$ probability. Repeating this would give a set of size $e^{O(d)}$ that works with high probability.

1.3 Related Work

In addition to the work of [AK03, GZ16] discussed above, several papers have looked at similar problems. When $\sigma = 0$, the problem becomes the standard one of Reed-Solomon decoding with relative Hamming distance $\rho$. Efficient algorithms such as Berlekamp-Massey [WB86] give unique decoding for all $\rho < 1/2$. For $\rho > 1/2$, while unique decoding is impossible, a polynomial size list decoding is possible—with sufficiently many samples—for all $\rho < 1$ [GS98].

Other work has looked at robust estimation for distributions. In the field of robust statistics (see, e.g., [Hub11]), as well as some recent papers in theoretical computer science (e.g. [LRV16, DKK+16, CSV16]), one would like to estimate properties of a distribution from samples that have $\rho$ chance of being adversarial. In some such cases, list decoding for $\rho > 1/2$ is possible [CSV16].

2 Preliminaries

For a function $f : [-1, 1] \to \mathbb{R}$ and interval $I \subseteq [-1, 1]$, we define the $\ell_q$ norm on the interval to be

$$\|f\|_{I,q} := \left( \int_I |f(x)|^q dx \right)^{1/q},$$

where $\|f\|_{I,\infty} := \sup_{x \in I} |f(x)|$. We also define the overall $\ell_q$ norm $\|f\|_q := \|f\|_{[-1,1],q}$. We will need the following consequence of a generalization due to Nevai [Nev79] of Bernstein’s inequality from $\ell_\infty$ to $\ell_q$. The proof of this lemma is in Appendix A.

**Lemma 2.1.** Let $p$ be a degree $d$ polynomial. Let $I_1, \ldots, I_m$ partition $[-1, 1]$ between the Chebyshev extrema $\cos \frac{\pi j}{m}$, for some $m \geq d$. Let $r : [-1, 1] \to \mathbb{R}$ be piecewise constant, so that for each $I_k$ there
exists an \( x_k^* \in I_k \) with \( r(x) = p(x_k^*) \) for all \( x \in I_k \). Then there exists a universal constant \( C \) such that, for any \( q \geq 1 \),

\[
\|p - r\|_q \leq C \frac{d}{m} \|p\|_q.
\]

We recall the definition of the Chebyshev polynomials \( T_d(x) \), which we will use extensively.

**Definition 2.2.** The Chebyshev polynomials of the first kind \( T_d(x) \) are defined by the following recurrence: \( T_0(x) = 1, T_1(x) = x \) and \( T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \). They have the property that \( T_d(\cos \theta) = \cos(d\theta) \) for all \( \theta \).

### 3 \( \ell_1 \) regression

**Lemma 3.1.** Suppose \( m \geq C_d \epsilon \) for a large enough constant \( C \). Then, for any set of samples \( x_1, \ldots, x_n \) with all \( S_i = \{ j \mid x_j \in I_i \} \) nonempty, and any polynomial \( p \) of degree \( d \),

\[
\sum_i \left| \frac{|I_i|}{|S_i|} \sum_j p(x_j) \right| = (1 \pm \epsilon) \|p\|_1.
\]

**Proof.** When all \( |S_i| = 1 \), this is a restatement of Lemma 2.1 for \( q = 1 \). Otherwise, the LHS is the expectation of the \( |S_i| = 1 \) case, when randomly drawing a single \( j \) in each \( S_i \).

We will show a more precise statement than Lemma 1.2, which allows for a weaker \( \ell_1 \) version of the \( \alpha \)-good requirement on the samples:

**Definition 3.2.** For a set of samples \( (x_1, y_1), \ldots, (x_n, y_n) \), and the Chebyshev partition \( I_1, \ldots, I_m \), define \( S_j = \{ i \mid x_i \in I_j \} \). We say that the samples are \( (\alpha, \sigma) \) close to a polynomial \( p \) in \( \ell_1 \) if, for

\[
e_j := \min_{S' \subset S_j, |S'| \leq (1-\alpha)|S_j|} \max_{j \in S'} |p(x_j) - y_j|,
\]

we have

\[
\sum_j |I_j| e_j \leq \sigma.
\]

If the samples are \( \alpha \)-good, then each \( e_j \leq \sigma \), so \( \sum |I_j| e_j \leq (\sum |I_j|)(\max_i e_j) \leq 2\sigma \) and the samples are \( (\alpha, 2\sigma) \) close to \( p \) in \( \ell_1 \).

We now state Lemma 3.3 which is a more precise statement of Lemma 1.2.

**Lemma 3.3.** Let \( \alpha < 1/2 \), and \( m \geq C_d \epsilon \) for a large enough constant \( C \) and some \( \epsilon \leq (1 - 2\alpha)/4 \). Then, given any samples \( (\alpha, \sigma) \) close to \( p \) in \( \ell_1 \), the degree \( d \) polynomial solution to the \( \ell_1 \) regression problem

\[
\hat{p} = \arg \min_q \sum_{j \in [m]} \left| \frac{|I_j|}{|S_j|} \sum_{i \in S_j} |y_i - q(x_i)| \right|
\]

has

\[
\|\hat{p} - p\|_1 \leq \frac{2\sigma}{1 - 2\alpha}.
\]

**Proof.** In each interval, we consider the \( \alpha|S_j| \) coordinates maximizing \( |y_i - p(x_i)| \) to be ‘bad’, and the rest to be ‘good’. Let the set of ‘bad’ and ‘good’ coordinates in \( S_j \) be denoted by \( B_j \) and \( G_j \).
respectively, and define \( \text{obj}(f) = \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \sum_{i \in S_j} |y_i - f(x_i)| \). By definition \( \text{obj}(\hat{p}) \leq \text{obj}(p) \). This gives us

\[
0 \geq \text{obj}(\hat{p}) - \text{obj}(p)
\]

\[
= \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left( \sum_{i \in S_j} |y_i - \hat{p}(x_i)| - \sum_{i \in S_j} |y_i - p(x_i)| \right)
\]

\[
= \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left( \left( \sum_{i \in G_j} |y_i - \hat{p}(x_i)| - |y_i - p(x_i)| \right) + \left( \sum_{i \in B_j} |y_i - \hat{p}(x_i)| - |y_i - p(x_i)| \right) \right).
\]

From the triangle inequality, we have \( |y_i - \hat{p}(x_i)| \geq |p(x_i) - \hat{p}(x_i)| - |y_i - p(x_i)| \) and \( |y_i - \hat{p}(x_i)| - |y_i - p(x_i)| \geq -|p(x_i) - \hat{p}(x_i)| \). This gives

\[
0 \geq \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left( \sum_{i \in G_j} |p(x_i) - \hat{p}(x_i)| - 2|y_i - p(x_i)| - \sum_{i \in B_j} |p(x_i) - \hat{p}(x_i)| \right)
\]

Now, recall that our samples are \( (\alpha, \delta) \) close to \( p \) in \( \ell_1 \). This means (by Definition 3.2) that for any ‘good’ \( i \), \( |y_i - p(x_i)| \leq \epsilon \) for a set of \( \epsilon \)'s that satisfy \( \sum |I_j| \epsilon_j \leq \sigma \). Therefore, for these \( \epsilon \),

\[
0 \geq \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left( \sum_{i \in G_j} |p(x_i) - \hat{p}(x_i)| - \sum_{i \in B_j} |p(x_i) - \hat{p}(x_i)| \right) - \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left( \sum_{i \in G_j} |y_i - p(x_i)| \right)
\]

\[
\geq \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left( \sum_{i \in G_j} |p(x_i) - \hat{p}(x_i)| - \sum_{i \in B_j} |p(x_i) - \hat{p}(x_i)| \right) - \sum_{j \in [m]} \frac{|I_j|}{|S_j|} ((1 - \alpha)|S_j| \epsilon_j)
\]

Using Lemma 3.1 on \( p - \hat{p} \) we now get

\[
0 \geq (1 - \alpha)(1 - \epsilon)\|p - \hat{p}\|_1 - \alpha(1 + \epsilon)\|p - \hat{p}\|_1 - (1 - \alpha)\sigma
\]

or

\[
(1 - 2\alpha - \epsilon)\|p - \hat{p}\|_1 \leq (1 - \alpha)\sigma,
\]

Hence

\[
\|p - \hat{p}\|_1 \leq \frac{(1 - \alpha)\sigma}{1 - 2\alpha - \epsilon}.
\]

For \( \epsilon \leq \frac{1 - 2\alpha}{2} \), this gives the desired

\[
\|p - \hat{p}\|_1 \leq \frac{2\sigma}{1 - 2\alpha}.
\]

\[
\]

4 \hspace{1cm} \ell_\infty \hspace{1cm} \text{regression}

Given a set of \( \alpha < \frac{1}{2} \)-good samples \( S \), our goal is to find a degree \( \hat{p} \) polynomial \( q \) with \( \|\hat{p} - p\|_\infty \leq (2 + \epsilon)\sigma \). We start by proving Lemma 1.3, which we restate below for clarity:
Algorithm 1 Refinement method, analyzed in Lemma 1.3 for $r = 0$

1: procedure $\text{Refine}(S = \{(x_i, y_i)\}, \hat{p})$
2: \quad $\tilde{y}_j \leftarrow \text{median}_{x_i \in I_j} y_i - \hat{p}(x_i)$
3: \quad Choose arbitrary $\tilde{x}_j \in I_j$
4: \quad Fit degree $d$ polynomial $r$ minimizing $\|r(\tilde{x}_j) - \tilde{y}_j\|_\infty$
5: \quad $\hat{p}' \leftarrow \hat{p} + r$
6: \quad return $\hat{p}'$
7: end procedure

Algorithm 2 Complete recovery procedure

1: procedure $\text{Approx}(S)$
2: \quad $\hat{p}^{(0)} \leftarrow \text{result of } \ell_1 \text{ regression}$
3: \quad for $i \in [0, O(\log_1/\epsilon) d]$ do
4: \quad \quad $\hat{p}^{(i+1)} \leftarrow \text{Refine}(S, \hat{p}^{(i)})$
5: \quad end for
6: \quad return $\hat{p}^{(O(\log_1/\epsilon) d)}$
7: end procedure

Lemma 1.3. Let $\epsilon, \alpha < \frac{1}{2}$, and suppose the set $S$ of samples is $\alpha$-good for the size $m = O(d/\epsilon)$ Chebyshev partition. Let $\tilde{x}_j \in I_j$ be chosen arbitrarily, and let $\tilde{y}_j = \text{median}_{x_i \in I_j} y_i$. Then the degree $d$ polynomial $\hat{p}$ minimizing

$$\max_{j \in [m]} |\hat{p}(\tilde{x}_j) - \tilde{y}_j|$$

satisfies $\|\hat{p} - p\|_\infty \leq (2 + \epsilon)\sigma + \epsilon\|p\|_\infty$.

Proof. Since more than half the points in any interval $I_j$ are such that $|y_j - p(x_j)| \leq \sigma$, and $p$ is continuous, there must exist an $x'_j \in I_j$ satisfying

$$|\tilde{y}_j - p(x'_j)| \leq \sigma. \quad (1)$$

We now define three piecewise-constant functions, $r(x), \hat{r}(x),$ and $\tilde{r}(x)$, to be such that within each interval $I_j$ we have $r(x) = p(x'_j)$, $\hat{r}(x) = \hat{p}(\tilde{x}_j)$, and $\tilde{r}(x) = \tilde{y}_j$. By Lemma 2.1, for our choice of $m$ we have

$$\|p - r\|_\infty \leq \epsilon\|p\|_\infty \quad \text{and} \quad \|\hat{p} - \hat{r}\|_\infty \leq \epsilon\|\hat{p}\|_\infty. \quad (2)$$

We also have by (1) that $\|r - \tilde{r}\|_\infty \leq \sigma$, so by the triangle inequality

$$\|p - \tilde{r}\|_\infty \leq \sigma + \epsilon\|p\|_\infty. \quad (3)$$

Now, our choice of $\hat{p}$ ensures that

$$\|\tilde{r} - \hat{r}\|_\infty = \max_{j \in [m]} |\hat{p}(\tilde{x}_j) - \tilde{y}_j| \leq \max_{j \in [m]} |p(\tilde{x}_j) - \tilde{y}_j| \leq \|p - \tilde{r}\|_\infty \leq \sigma + \epsilon\|p\|_\infty.$$

Combining with (2) and (3) gives by the triangle inequality that

$$\|\hat{p} - p\|_\infty \leq 2\sigma + 2\epsilon\|p\|_\infty + \epsilon\|\hat{p}\|_\infty. \quad (4)$$
To finish the proof, we just need a bound on $\|\tilde{p}\|_\infty$. Note that (4) implies
\[ \|\tilde{p}\|_\infty \leq 2\sigma + (1 + 2\epsilon)\|p\|_\infty + \epsilon\|\tilde{p}\|_\infty, \]
and since $\epsilon \leq 1/2$, this implies
\[ \|\tilde{p}\|_\infty \leq 4\sigma + (2 + 4\epsilon)\|p\|_\infty. \]
Plugging into (4) and rescaling $\epsilon$ down by a constant factor gives the result. 

Before we prove Theorem 1.4, we briefly describe the algorithm. Algorithm 2 first sets its initial estimate to the polynomial produced by $\ell_1$ regression. It then continues to refine the estimate it has by using Algorithm 1 for $O(\log_{1/\epsilon} d)$ iterations.

**Theorem 1.4.** Let $\epsilon, \alpha < \frac{1}{2}$, and suppose the set $S$ of samples is $\alpha$-good for the size $m = O(d/\epsilon)$ Chebyshev partition. The result $\tilde{p}$ of Algorithm 2 is a degree $d$ polynomial satisfying $\|\tilde{p} - p\|_\infty \leq (2 + \epsilon)\sigma$. Its running time is that of solving $O(\log_{1/\epsilon} d)$ linear programs, each with $O(d)$ variables and $O(|S|)$ constraints.

**Proof.** Our algorithm proceeds by iteratively improving a polynomial approximation $\tilde{p}$ to $p$, so that $\|p - \tilde{p}\|_\infty$ improves at each stage. Our notation will be borrowed from the notation in the statements of Algorithms 2 and 1. Let $\tilde{p}^{(t)}(x)$ be the $t^{th}$ estimate of $p$ found by Algorithm 2 and let $e_t(x) = (p - \tilde{p}^{(t)})(x)$ be the error of the $t^{th}$ estimate $\tilde{p}^{(t)}$. $r_t$ is the polynomial $r$ found by the $t^{th}$ iteration of Algorithm 1, i.e. $r_t = \arg\min_r \|r(x) - y\|_\infty$ where $y = \text{median}_{x \in I_t} y_t - \tilde{p}^{(t)}(x_t)$, and the minimum is taken over all degree $d$ polynomials. Lemma 1.3 now implies
\[ \|r_t(x) - e_t(x)\|_\infty \leq (2 + \epsilon)\sigma + \epsilon\|e_t(x)\|_\infty. \]
Observe that $r_t(x) - e_t(x) = (\tilde{p}^{(t+1)}(x) - \tilde{p}^{(t)}(x)) - (p(x) - \tilde{p}^{(t)}(x)) = -e_{t+1}(x)$, which gives us
\[ \|e_{t+1}(x)\|_\infty \leq (2 + \epsilon)\sigma + \epsilon\|e_t(x)\|_\infty. \]
Proceeding by induction and using the geometric series formula we see
\[ \|p - \tilde{p}^{(t)}\|_\infty \leq \frac{2 + \epsilon}{1 - \epsilon} \sigma + \epsilon^t\|e_0\|_\infty \leq (2 + 4\epsilon)\sigma + \epsilon^t\|e_0\|_\infty. \]
Rescaling $\epsilon$, we see that in a number of iterations logarithmic in the quality of our initial solution, we find a $\tilde{p}$ such that $\|\tilde{p} - p\|_\infty \leq (2 + \epsilon)\sigma$.

Finally, we analyze the quality of the initial solution produced by $\ell_1$ regression. By Lemma 1.2, $\|e_0\|_1 \leq O(\sigma)$. Applying the Markov brothers’ inequality to the degree $d + 1$ polynomial $Q(x) = \int_{-1}^{x} e_0(u)du$, we get
\[ \|e_0\|_\infty = \max_{x \in [-1,1]} |Q'(x)| \leq (d + 1)^2 \max_{x \in [-1,1]} |Q(x)| \leq (d + 1)^2\|e_0\|_1 \leq O(d^2\sigma). \]
Hence the number of iterations required is $O(\log_{1/\epsilon} d)$. 

**Proof.** It is enough to show that for $m = O(d)$, drawing $O(\frac{d}{\epsilon} \log \frac{d}{\alpha})$ samples from the Chebyshev distribution or $O(\frac{d}{\epsilon} \log (\frac{1}{\delta}))$ samples from the uniform distribution gives us an $\alpha$-good sample for some $\alpha < \frac{1}{2}$ with probability $1 - \delta$. The corollary then follows from Theorem 1.4.
Let \( k \) be the total number of samples taken, and let \( p_j \) denote the probability that any sampled \( x_i \) lies in the \( j \)th interval \( I_j \). With Chebyshev sampling,

\[
p_j = \int_{\cos\left(\frac{\pi(j+1)}{m}\right)}^{\cos\left(\frac{\pi}{m}\right)} \frac{1}{\sqrt{1-x^2}} = \arcsin\left(\cos\left(\frac{\pi(j+1)}{m}\right)\right) - \arcsin\left(\cos\left(\frac{\pi}{m}\right)\right) = \frac{1}{m}
\]

for all \( j \). With uniform sampling, \( p_j = \Theta\left(\min\left(j,m+1-j\right)\right)\).

Let \( X_j \) be the number of samples that appear in \( I_j \), and \( Y_j \) be the number of these samples that are outliers. We have \( \mathbb{E}[X_j] = kp_j \), so by a Chernoff bound,

\[
\Pr[X_j \leq \frac{1}{2}kp_j] \leq e^{-\Omega(kp_j)}.
\]

(5)

The outliers are then chosen independently, with expectation \( \rho X_j \), so

\[
\Pr[Y_j \geq \alpha X_j \mid X_j] \leq e^{-\Omega((\alpha-\rho)^2X_j)}.
\]

(6)

Setting \( \alpha = \rho + \frac{1}{2} \), we have that conditioned on (5) not occurring, (6) occurs with at most \( e^{-\Omega(kp_j)} \) probability. If neither occur, then less than an \( \alpha \) fraction of the samples in \( I_j \) are outliers. Hence, by a union bound over the intervals, the samples are \( \alpha \)-good with probability at least

\[
1 - 2\sum_{i=1}^{m} e^{-\Omega(kp_j)}.
\]

(7)

In the Chebyshev setting, we have \( p_j = 1/m \) and \( k = O(m \log(m/\delta)) \), making (7) at least \( 1 - 2\delta \) for appropriately chosen constants. In the uniform setting, we similarly have

\[
\sum_{j=1}^{m} e^{-\Omega(kp_j)} = \sum_{j=1}^{m} e^{-\Omega(m^2 \log(1/\delta) \cdot \min(j,1+m-1)/m^2)} \leq 2\sum_{j=1}^{m/2} \delta^j \leq 3\delta,
\]

making (7) at least \( 1 - 6\delta \). Rescaling \( \delta \) gives the result, that the samples are \( \alpha \)-good for some \( \alpha < 1/2 \) with probability at least \( 1 - \delta \).

\[\square\]

5 Impossibility results

5.1 Sample Complexity

**Lemma 5.1.** If the \( x_i \) are sampled uniformly then it is not possible to get an \( O(1) \) approximation in \( \ell_\infty \) norm to the original function in \( o(d^2) \) samples and 1/4 failure probability.

**Proof.** Consider an algorithm that gives a \( C \)-approximation given \( s \) uniform samples with noise level \( \sigma = 1 \) and zero outliers, for \( C = O(1) \) and \( s = o(d^2) \). We will construct two polynomials with \( \ell_\infty \) distance more than 2\( C \), but for which the samples have a constant chance of being indistinguishable.

Define the polynomials \( g(x) = 0 \) and \( f(x) := T_d \left( x + \frac{\alpha}{d^2} \right) \) where \( T_d \) is the degree \( d \) Chebyshev polynomial of the first kind, for some \( d \geq 4 \) and constant \( \alpha = 4\sqrt{2(C-1)} \). By construction, \( |f(x)| \leq 1 \) for \( x \in [-1,1-\frac{\alpha}{d^2}] \), but \( |f(1)| > 2C \) because for \( d \geq 4 \)

\[
|f(1)| = \left| T_d \left( 1 + \frac{\alpha}{d^2} \right) \right| = \left| \cosh\left( d \arccosh\left( 1 + \frac{\alpha}{d^2} \right) \right) \right| > 2C.
\]
Lemma 5.4. For any distribution on sets the set of degree-\(d\) polynomials
\[
\text{L}_H
\]
Hence the probability that at least one \(\{x \in \mathbb{R}^d \mid b < x \leq a\}\) contains these \(\rho = \Omega(1)\).

Suppose one always observes samples of the form \((x_i, 0)\). These samples are within \(\sigma = 1\) of both \(f(x)\) and \(g(x)\) if they lie in the region \([-1, 1 - \frac{\sqrt{2}}{d}]\). The chance that all samples \(x_i\) lie in this region is at least \((1 - \frac{\sigma}{2d})^s \geq e^{-\sigma_0/d^2}\), which is \(e^{-o(1)} \geq 1/2\). Hence there is at least a 1/2 chance that the samples from \(f\) and \(g\) are indistinguishable, so the algorithm has a failure probability more than 1/4.

Our next lower bound will use the following lemmas, proven in Appendix B.

Lemma 5.2. Let \(d \geq 1\). For any point \(b \in [-1, 1]\), there exists a degree \(d\) polynomial \(p_b\) such that \(p_b(b) = 1, \|p_b\|_\infty = 1\), and
\[
|p_b(x)| \leq \frac{2}{d|x - b|}
\]
for all \(x \in [-1, 1]\).

Lemma 5.3. For any \(d\) and \(\alpha > 0\), let \(m = d\sqrt{\alpha}/2\), and define \(b_j = -1 + \frac{2}{m}j\) for \(j \in [m]\). Consider the set of degree-\(d\) polynomials
\[
f_S(x) = \sum_{j \in S} p_{b_j}^2(x)
\]
for \(S \subseteq [m]\). For any \(x \in [-1, 1]\), let \(k_x \in [m]\) minimize \(|b_{k_x} - x|\). Then for any \(S \subseteq [m]\),
\[
f_{\{k_x\} \cap S}(x) \leq f_S(x) \leq f_{\{k_x\} \cap S}(x) + \alpha.
\]

Lemma 5.4. For any distribution on sets \(x = (x_1, \ldots, x_s)\) of \(s = o(d \log d)\) sample points with independent outlier chance \(\rho = \Omega(1)\), it is not possible to get a robust \(O(1)\) approximation in \(\ell_\infty\) norm to the original function with 1/4 failure probability.

Proof. Let \(m = d/\sqrt{12C}\), and \(f_S\) and \(k_x\) be as in Lemma 5.3 for \(\alpha = \frac{1}{3C}\). For any \(x\), let \(L_j := \{i \in [s] \mid k_{x_i} = j\}\). Since these sets are disjoint, there must exist at least \(m/2\) different \(j\) for which \(|L_j| \leq 2s/m = o(\log d)\). Let \(B \subseteq [m]\) contain these \(j\). We say that a given \(L_j\) is “outlier-full” if all of the \(x_i \in L_j\) are outliers, which for \(j \in B\) happens with probability at least
\[
\rho^{|L_j|} \geq 2^{-o(\log d)} \geq 1/\sqrt{d}.
\]
Hence the probability that at least one \(L_j\) is outlier-full is at least
\[
1 - (1 - 1/\sqrt{d})^{|B|} \geq 1 - e^{-m/(2\sqrt{d})} > 0.99.
\]
Suppose the true polynomial $p$ to be learned is $f_S$ for a uniformly random $S$, and consider the following adversary. At least one $L_j$ is outlier-full, she arbitrarily picks one such $j^*$ and sets $S'$ to the symmetric difference of $S$ and $\{j^*\}$. She then flips a coin, and with 50% probability outputs $(x_i, f_S(x_i))$ for each $i$, and otherwise outputs $(x_i, f_{S'}(x_i))$ for each $i$. This is valid for $\sigma = \frac{1}{3\ell}$, because for each $i \in [s]$, either $k_{x_i} = j^*$ (in which case $x_i \in L_{j^*}$ is an outlier) or $\{k_{x_i}\} \cap S = \{k_{x_i}\} \cap S'$ (in which case Lemma 5.5 implies $|f_S(x_i) - f_{S'}(x_i)| \leq \alpha = \frac{1}{3\ell}$).

Because the distribution on $j^*$ is independent of $S$, the distribution of $S'$ is also uniform, so the algorithm cannot distinguish whether it received $f_S$ or $f_{S'}$. But $\|f_S - f_{S'}\|_\infty = \|f_{j^*}\|_\infty = 1 > 2C\sigma$, so the algorithm’s output cannot satisfy both cases simultaneously. Hence, in the 99% of cases where one $L_j$ is outlier-full, the algorithm will have 50% failure probability, for $49.5% > 1/4$ overall.

### 5.2 Approximation Factor

Any algorithm that relies on the result of the $\ell_\infty$ projection cannot do significantly better. There are two functions $p(x)$ and $f(x)$ such that $|p(x) - f(x)| \leq \sigma$ for all $x \in [-1, 1]$, however the $\ell_\infty$ projection to the space of all degree 1 polynomials is almost $2\sigma$ away in $\ell_\infty$ norm from $p$.

**Lemma 5.5.** Let $d = 1$, $p(x) = \sigma$ be a constant function and $f(x) = \frac{2}{\alpha} \max (0, x - (1 - 2\alpha))$ where $\alpha < \frac{1}{2}$. Note that $|p(x) - f(x)| \leq \sigma$ for all $x \in [-1, 1]$. The result $q = \arg\min_r \|f - r\|_\infty$ of $\ell_\infty$ projection of $f$ to the space of degree-1 polynomials satisfies $\|p(x) - q(x)\|_{[-1,1],\infty} \geq (2 - \alpha)\sigma$.

**Proof.** Observe that $f(x) = 0$ for $x \in [-1, (1 - 2\alpha)]$ and $f(x) = 2 \left(\frac{x}{\alpha} - 1\right) + \sigma$ in $[1 - 2\alpha, 1]$. The maximum difference between two linear equations in a closed interval is attained at the endpoints of that interval. This tells us

$$q(x) = \arg\min_r \max\{|f(-1) - r(-1)|, |f(1 - 2\alpha) - r(1 - 2\alpha)|, |f(1) - r(1)|\}$$

$$= \arg\min_r \max\{|r(-1)|, |r(1 - 2\alpha)|, |2\sigma - r(1)|\}.$$

Let $q(x) = ax + b$. $q(x)$ will be such that $f(-1) > q(-1)$, $f(1 - 2\alpha) < q(1 - 2\alpha)$ and $f(1) > q(1)$ (see Figure 4), and so we want

$$\arg\min_{(a,b)} \max\{a - b, a + b - 2\alpha a, 2\sigma - (a + b)\}.$$

This function is minimized when all three terms are equal, which happens when $a = \sigma$ and $b = -\alpha\sigma$. This gives $q(x) = 2\sigma x - \alpha\sigma$, and $\|q(x) - p(x)\|_{[-1,1],\infty} = (2 - \alpha)\sigma$.

We now show that that one cannot hope for a proper $(1 + \epsilon)$-approximate algorithm, even with no outliers. We will present a set of polynomials such that no two are more than 2-apart, but for which no single polynomial lies within $\alpha > 1$ of all polynomials in the set. Then an adversary with $\sigma = 1$ can output a function $f(x)$ independent of the choice of polynomial in the set, forcing the algorithm to by $\alpha$-far when recovering some polynomial in the set.

**Lemma 5.6.** There exist three degree $\leq 2$ polynomials, all within 2 of each other over $[-1, 1]$, such that any single quadratic function has $\ell_\infty$ distance more than 1.09 from one of the three.

**Proof.** Consider the polynomials $p_1(x) = (x + 1)$, $p_2(x) = (1 - x)$, $p_3(x) = \frac{3 + 2\sqrt{2}}{2} (1 - x^2)$, and let $v$ denote $\frac{1}{3 + 2\sqrt{2}}$ (see Figure 5). Observe that $p_1$ and $p_2$ are at distance 2 from each other at $x = 1, -1$. Also $p_1$ is at distance 2 from $p_3$ at $x = -v$, and similarly $p_2$ is at distance 2 from $p_3$. 

\[\square\]
Figure 4: $q(x)$ makes an error of $(2 - \alpha)\sigma$ with $p(x)$.

Figure 5: The quadratic that passes through the four points above does not satisfy the inequality $ax^2 + c \leq p_1(x) + 1$ in the range $[-1, -v]$, for $v = \frac{1}{3+2\sqrt{2}}$.

at $x = v$. Hence, any polynomial that wants to 1-approximate all the $p_i$’s necessarily has to go through the points $(1, 1), (-1, 1), (v, 2 - v), (-v, 2 - v)$. By symmetry, we know that any quadratic that goes through these will be of the form $y = ax^2 + c$. Substituting these values in the equation and solving for $a$ and $c$ we see that $a = \frac{1}{v+1}$ and $c = 1 - \frac{1}{v+1}$.

Since the quadratic $ax^2 + c$ has to 1-approximate the $p_i$ it must be the case that $ax^2 + c \leq p_1(x) + 1 = x + 2$ at all points. However this inequality is not satisfied in the interval $(-1, -v)$ as shown in Figure 5. This is because $p_1(x) + 1$ is the line between two points on the curve $ax^2 + c$, which is a concave function for $a < 0$. Running a program to optimize parameters, we see that the best approximation to these polynomials will make an error of at least 1.09 with one of the polynomials (see Figure 2).

\[\Box\]

**Lemma 5.7.** There exist $O(d)$ degree 2 polynomials such that any degree $d$ polynomial that tries to approximate all these polynomials has to make an error of $\Omega(\frac{1}{d^{1/2}})$ with at least one of the degree 2 polynomials.

**Proof.** We will consider a set of quadratic equations such that any two are at most distance $\sigma$ from each other, and such that any polynomial that has to $\sigma$-approximate them needs to perform $2d$ oscillations. As no degree $d$ polynomial can perform more than $d$ oscillations, there is at least one
Figure 6: Construction for Lemma 5.7. On the left, any 1-approximator necessarily has to go through the three points. On the right, placing translated copies of these functions force any 1-approximation to perform more than \(d\) oscillations. These figures are not to scale.

Figure 7: One element of \(f_S\) from Lemma 5.3.

oscillation that this polynomial does not perform, and hence the approximating polynomial will make an error of at least \(1 + h\) where \(h\) is the height of the oscillation.

Observe that \(\| (1 - Ax^2) - A(x + c)^2 \|_\infty = 1 - \frac{A^2}{2} \) and this is achieved at \(x = \pm \frac{c}{2}\). This means the set \(S_0 = \{1 - Ax^2, A(x - c)^2, A(x + c)^2, Ax^2 + \frac{Ac^2}{2} \}\) has every polynomial at most at distance \(2\sigma := 1 - \frac{A^2}{2}\) from each other. Any \(\sigma\) approximator, hence, necessarily has to go through the three points as shown in Figure 6. Define \(S_t = \{1 - A(x - 2ct)^2, A(x - 2ct - c)^2, A(x - 2ct + c)^2, A(x - 2ct)^2 + \frac{Ac^2}{2} \}\) Now observe that any \(\sigma\) approximator to \(S = \bigcup_{t \in [-1.5d, 1.5d]} S_t\) will necessarily have to perform \(2d\) oscillations. We will now define the parameters such that these \(2d\) oscillations take place in the range \([-1, 1]\) and every pair of functions is at distance at most \(1 - \frac{Ac^2}{2}\) from each other. Set \(A = \frac{1}{2d}\) and \(c = \frac{1}{4d}\). This ensures that every pair of elements in \(S_t\) are at most at distance \(1 - \frac{1}{16d^3}\) from each other.

We now show that if \(p \in S_t\) and \(q \in S_{t'}\) for \(t < t'\), then \(\| p - q \|_\infty \leq 1 - \frac{1}{64d^3}\). Observe that it is enough to check this for \(p = 1 - A(x - 2ct)^2\) and \(q = A(x - 2ct' + c)^2\). Because of our choice of \(A, c\)

\[
|1 - A(x - 2ct)^2 - A(x - 2ct' + c)^2| = \left| 1 - \frac{1}{2d} \left( \left( x - \frac{2t}{4d} \right)^2 + \left( x - \frac{2t' - 1}{4d} \right)^2 \right) \right| \\
\leq \left| 1 - \frac{1}{2d} \left( \frac{2}{4} \cdot \frac{(2(t' - t) - 1)^2}{16d^2} \right) \right| \\
\leq 1 - \frac{1}{64d^3}
\]

Finally, observe that we force the approximating polynomial to perform one oscillation for every \(S_t\), and if \(c = \frac{1}{4d}\) the polynomial has to perform \(\frac{2d}{3}\) oscillations to \(\sigma\)-approximate every polynomial in \(S\) in the interval \([-1, 1]\) because it has to perform one oscillation in every interval of length \(3c\). Since no degree \(d\) polynomial can perform \(2d\) oscillations, there is at least one oscillation that it cannot
perform, and so the approximating polynomial necessarily has to make an error of at least $1 + h$ with one of the polynomials in $S$, where $h$ is the height of the oscillation, which is $\Omega(\frac{1}{d^3})$.

5.3 List decoding

We now show that if the probability of getting a bad sample were greater than $\frac{1}{2}$, then it is not possible to find $\text{poly}(d)$ polynomials of degree $O(d)$ such that one of the polynomials is close to the original polynomial.

**Theorem 5.8.** Consider any algorithm for robust polynomial regression that returns a set $L$ of polynomials from samples with outlier chance $\rho = 1/2$, such that at least one element of $L$ is a $C$-approximation to the true answer with $3/4$ probability. Then $\mathbb{E}[|L|] \geq \frac{3}{4} 2^{\Omega(d/\sqrt{C})}$.

**Proof.** Note that we may assume $d/C$ is larger than a sufficiently large constant, since otherwise the result follows from the fact that $|L| \geq 1$ whenever the algorithm succeeds. Let us then set $m = d/\sqrt{12C}$, and take $f_S$ from Lemma 5.3 for $\alpha = 1/(3C)$. For any $S \subseteq [m]$ and $x \in [-1, 1]$, the lemma implies either $f_\{\{\}(x) \leq f_S(x) \leq f_{\{\}(x) + \alpha$ or $f_{[m]}(x) - \alpha \leq f_S(x) \leq f_{[m]}(x)$.

Therefore, when observing any polynomial $f_S$ for $S \subseteq [m]$, the adversary for $\sigma = \alpha$ can ensure that any sample point $x$ is observed as $f_\{\{\}(x)$ with 50% probability, and $f_{[m]}(x)$ with 50% probability. This is because $p(x)$ is always within $\alpha$ of one of these, so the adversary can output that one if $x$ is an inlier, and the other one if $x$ is an outlier. For this adversary, the algorithm’s input is independent of the choice of $f_S$. Hence its distribution on output $L$ is also independent of $f_S$. But each $\hat{p} \in L$ can only be a $C$-approximation to at most one $f_S$. Hence, if $S \subseteq [m]$ is chosen at random, $\Pr[\exists \hat{p} \in L : \|\hat{p} - f_S\| \leq C\alpha] \leq \frac{\mathbb{E}[|L|]}{2^m}$.

This implies the result. □

**Acknowledgements**

We would like to thank user111 on MathOverflow [use] for pointing us to [Nev79].

**References**

[Ach13] Naum I Achieser. *Theory of approximation*. Courier Corporation, 2013.

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

[CSV16] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. *arXiv preprint arXiv:1611.02315*, 2016.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.
A Proof of Lemma 2.1

To prove Lemma 2.1, we need a a generalization of Bernstein’s inequality from $\ell_\infty$ to $\ell_q$ for all $q > 0$, which appears as Theorem 5 of [Nev79]. In the univariate case and setting its parameters $\gamma, \Gamma = 0$, it states

**Lemma A.1** (Special case of Theorem 5 of [Nev79]). There exists a universal constant $C$ such that, for any degree $d$ polynomial $p(x)$ and any $q > 0$,

$$
\int_{-1}^{1} |\sqrt{1-x^2}p'(x)|^q dx \leq C d^q \int_{-1}^{1} |p(x)|^q dx.
$$
Lemma 2.1. Let $p$ be a degree $d$ polynomial. Let $I_1, \ldots, I_m$ partition $[-1, 1]$ between the Chebyshev extrema $\cos \frac{\pi}{m}$, for some $m \geq d$. Let $r : [-1, 1] \to \mathbb{R}$ be piecewise constant, so that for each $I_k$ there exists an $x^*_k \in I_k$ with $r(x) = p(x^*_k)$ for all $x \in I_k$. Then there exists a universal constant $C$ such that, for any $q \geq 1$,

$$\|p - r\|_q \leq C \frac{d}{m} \|p\|_q.$$  

Proof. For any individual $I_k$ and $x^*_k \in I_k$, we have by Hölder’s inequality that

$$\int_{x \in I_k} |p(x) - p(x^*_k)|^q dx = \int_{x \in I_k} \left| \int_{x^*_k}^x p'(y) dy \right|^q dx \leq \int_{x \in I_k} |x - x^*_k|^{q-1} \int_{x^*_k}^x |p'(y)|^q dy dx \leq |I_k|^q \int_{x \in I_k} |p'(x)|^q dx.$$

Hence

$$\int_{x \in I_k} |p(x) - r(x)|^q dx \leq |I_k|^q \int_{x \in I_k} |p'(x)|^q dx. \quad (8)$$

We first consider $I_2, \ldots, I_{m/2}$, then separately consider the first interval $I_1$. The statement holds for the intervals $I_{m/2}, \ldots, I_m$ by symmetry. For each interval $I_k$ with $2 \leq k \leq \frac{m}{2}$, we have for all $x \in I_k$

$$|I_k| = \cos \left( \frac{k\pi}{m} \right) - \cos \left( \frac{(k+1)\pi}{m} \right) \leq \frac{k}{m^2} \leq \frac{\sin(k\pi/m)}{m} \lesssim \sqrt{1 - x^2},$$

where we use the notation $a \lesssim b$ to denote that there exists a universal constant $C$ such that $a \leq Cb$. Hence

$$\int_{x \in I_k} |p(x) - r(x)|^q dx \leq |I_k|^q \int_{x \in I_k} |p'(x)|^q dx \leq \frac{1}{(cm)^q} \int_{x \in I_k} \sqrt{1 - x^2} |p'(x)|^q dx$$

and so by Lemma A.1, for some constant $c$,

$$\int_{x \in I_2 \cup \cdots \cup I_{m/2}} |p(x) - r(x)|^q dx \leq \frac{1}{(cm)^q} \int_{-1}^1 \sqrt{1 - x^2} |p'(x)|^q dx \leq \left( \frac{d}{m} \|p\|_q \right)^q. \quad (9)$$

Now we consider the end $I_1$. By the Markov brothers’ inequality [Ach13],

$$\|p'\|_\infty \leq d^2 \|p\|_\infty.$$  

Let $x^* \in [-1, 1]$ such that $|p(x^*)| = \|p\|_\infty$, and let $I' = \{ y \in [-1, 1] \mid |x^* - y| \leq \frac{1}{2d^2} \}$. We have $p(y) \geq \|p\|_\infty / 2$ for all $y \in I'$, so

$$\|p\|_q^q \geq \|I'| (\|p\|_\infty / 2)^q \geq \frac{\|p\|_\infty^q}{2d^2 2^q}.$$  

Hence by (8), and using that $|I_1| = 1 - \cos \frac{\pi}{m} = \Theta\left( \frac{1}{m^2} \right)$,

$$\int_{x \in I_1} |p(x) - r(x)|^q dx \leq |I_1|^{q+1} \|p'\|_\infty^q \leq \frac{d^2 q}{(cm)^{2q+2}} \|p\|_\infty^q \lesssim \left( \frac{d \sqrt{2}}{cm} \right)^{2q+2} \|p\|_q^q.$$
For some constant $c$. The same holds for intervals $I_{m/2}, \ldots, I_m$ by symmetry, so combining with (9) gives
\[
\int_{-1}^{1} |p(x) - r(x)|^q dx \lesssim \left( \frac{d}{cm} \right)^q \left( 1 + 2^{q+1} \left( \frac{d}{m} \right)^2 \right) \|p\|^q_q
\]
or
\[
\|p - r\|_q \lesssim \frac{d}{m} \left( 1 + \left( \frac{d}{m} \right)^{2/q} \right) \|p\|_q.
\]
When $m \geq d$, the first term dominates giving the result. \hfill \Box

B Proof of Lemma 5.2

Lemma 5.2. Let $d \geq 1$. For any point $b \in [-1, 1]$, there exists a degree $d$ polynomial $p_b$ such that $p_b(b) = 1$, $\|p_b\|_\infty = 1$, and
\[
|p_b(x)| \leq \frac{2}{d|x-b|}
\]
for all $x \in [-1, 1]$.

Proof. We will show a stronger form of this lemma for even $d$ and $b = 0$, giving
\[
|p_0(x)| \leq \frac{1}{(d+1)|x|}.
\] (10)

By subtracting 1 from odd $d$, this implies the same for general $d$ with a $1/d$ rather than $1/(d+1)$ term. Then for general $b$ we take $p_b(x) = p_0((x - b)/2)$, which satisfies the lemma. So it suffices to show (10) for even $d$ and $b = 0$.

We choose $p_0(x)$ to be
\[
p(x) := (-1)^{d/2} \frac{T_{d+1}(x)}{(d+1)x}
\]
so that
\[
p(\cos \theta) = (-1)^{d/2} \frac{\cos((d+1)\theta)}{(d+1)\cos \theta}
\]
or, replacing $\theta$ with $\frac{\pi}{2} - \theta$ and using that $\sin(d\frac{\pi}{2} + \psi) = (-1)^{d/2} \sin \psi$,
\[
p(\sin \theta) = \frac{\sin((d+1)\theta)}{(d+1)\sin \theta}.
\]
Since $|\sin((d+1)\theta)| \leq 1$, this immediately gives (10); we just need to show $p(0) = \|p\|_\infty = 1$. By L'Hôpital’s rule, we have
\[
p(0) = \frac{(d+1)\cos((d+1)\cdot 0)}{(d+1)\cos 0} = 1.
\]
If $\theta \geq 1.1/(d+1)$, then $|\sin \theta| \geq 1/(d+1)$, and so (10) implies $|p(\sin \theta)| \leq 1$. Since $p$ is symmetric, all that remains is to show $|p(\sin \theta)| \leq 1$ for $0 < \theta < 1.1/(d+1)$.

The maximum value of $|p(\sin \theta)|$ will either appear at an endpoint of this interval—which we have already shown is at most 1—or at a zero of the derivative. We have
\[
\frac{\partial}{\partial \theta} p(\sin \theta) = \frac{(d+1)\cos((d+1)\theta)}{(d+1)\sin \theta} - \frac{\sin((d+1)\theta)}{(d+1)\sin \theta}^2 \cdot (d+1)\cos \theta
\]
\[
= \frac{(d+1)\cos((d+1)\theta)\sin \theta - \sin((d+1)\theta)\cos \theta}{(d+1)\sin^2 \theta}.
\]
For all $0 < \psi$, we have the inequalities $\psi - \psi^3/6 < \sin \psi < \psi$ and $1 - \psi^2/2 < \cos \psi < 1 - \psi^2/2 + \psi^4/24$. Hence for $0 < \theta < 1.1/(d+1)$, the denominator is positive and the numerator is less than

\[
(d+1)(1 - (d+1)^2\theta^2/2 + (d+1)^4\theta^4/24)\theta - ((d+1)\theta - (d+1)^3\theta^3/6)(1 - \theta^2/2)
\]

\[
= \theta^3(-((d+1)^3/2 + (d+1)/2 + (d+1)^3/6) + \theta^5((d+1)^5/24 - (d+1)^3/12))
\]

\[
\leq -\frac{5}{18}(\theta(d+1))^3 + (\theta(d+1))^5/24
\]

\[
\leq -0.22(\theta(d+1))^3 < 0.
\]

Thus $p(\sin \theta)$ does not have a local maximum on $(0, 1.1/(d+1))$, so $\|p\|_\infty \leq 1$, finishing the proof.

**Lemma 5.3.** For any $d$ and $\alpha > 0$, let $m = d\sqrt{\alpha}/2$, and define $b_j = -1 + 2m^2 j$ for $j \in [m]$. Consider the set of degree-$d$ polynomials

\[
f_S(x) = \sum_{j \in S} p_{b_j}^2(x)
\]

for $S \subseteq [m]$. For any $x \in [-1, 1]$, let $k_x \in [m]$ minimize $|b_{k_x} - x|$. Then for any $S \subseteq [m]$,

\[
f_{\{k_x\} \cap S}(x) \leq f_S(x) \leq f_{\{k_x\} \cap S}(x) + \alpha.
\]

**Proof.** Since $f_{S \cup \{k_x\}}(x) = f_{\{k_x\}}(x) + f_{S \setminus \{k_x\}}(x)$ for any $S$, it is sufficient to prove the result when $k_x \notin S$. Then

\[
f_S(x) \leq \sum_{j \neq k_x} p_{b_j}^2(x) \leq \sum_{j \neq k_x} \frac{4}{d^2|x - b_j|^2} \leq \sum_{j \neq k_x} \frac{4}{d^2((2|j - k_x| - 1)/m)^2}
\]

\[
< \frac{2m^2}{d^2} \sum_{j=1}^\infty \frac{1}{(2j-1)^2} < \frac{2m^2 \pi^2}{d^2 \cdot 6}
\]

\[
< \alpha
\]

as desired.