58th Annual IEEE Symposium on Foundations of Computer Science

Robust polynomial regression up to the information theoretic limit

Daniel Kane
Department of Computer Science and Engineering / Department of Mathematics
University of California, San Diego
La Jolla, CA, USA
dakane@ucsd.edu

Sushrut Karmalkar, Eric Price
Department of Computer Science
The University of Texas at Austin
Austin, TX, USA
{*sushrutk, ecprice*}*@cs.utexas.edu*

***Abstract*—We consider the problem of *robust polynomial regression*, where one receives samples that are usually within a small additive error of a target polynomial, but have a chance of being arbitrary adversarial outliers. Previously, it was known how to efficiently estimate the target polynomial only when the outlier probability was subconstant in the degree of the target polynomial. We give an algorithm that works for the entire feasible range of outlier probabilities, 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.**

***Keywords*-polynomial regression, robust regression, robust recovery, learning, approximation;**

I. 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 [1], including computer graphics [2], machine learning [3], and statistics [4]. 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 [5]. 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 χ (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 ρ; if the sample is not an outlier, then $|w_i| \leq \sigma$ for some σ. Other than the random choice of outlier locations, the noise is adversarial. One would like to recover $\widehat{p}$ with $\|p - \widehat{p}\|_\infty \leq C\sigma$ using as few samples as possible, with high probability.

In other contexts an ℓ_∞ 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 ℓ_∞ requirement on the input allows for the strong ℓ_∞ bound on the output.

Arora and Khot [5] 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 [6] 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 [7] showed how to solve the problem in polynomial time for ρ 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 ($\|p - \widehat{p}\|_\infty \lesssim \sigma \left(1 + \frac{\|p\|_\infty}{\sigma}\right)^{0.01}$). 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 σ; 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.

A. Algorithmic results

The problem formulation has randomly placed outliers, which is needed to ensure that we can estimate

DOI 10.1109/FOCS.2017.43

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 I.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 "*α*-good" for the partition if, in every interval* I_j*, strictly less than an* α *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 α- 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 ℓ_1 regression *almost* solves the problem: it satisfies all the requirements, except that the resulting error $\|\widehat{p} - p\|$ is bounded in ℓ_1 not ℓ_∞:

Lemma I.2. *Suppose the set* S *of samples is* α*-good for the size* $m = O(d)$ *Chebyshev partition, for constant* $\alpha < \frac{1}{2}$*. Then the result* $\widehat{p}$ *of* ℓ_1 *regression*

$$\underset{\substack{\text{degree-}d\\ \text{polynomial } \widehat{p}}}{\arg\min} \sum_{i=1}^{n} |I_j| \operatorname*{mean}_{x_i \in I_j} |y_i - \widehat{p}(x_i)|$$

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

This has a nice parallel to the $d = 1$ case, where ℓ_1 regression is the canonical robust estimator.

As shown by the Chebyshev polynomial in Figure 1, ℓ_1 regression might not give a good solution to the ℓ_∞ problem. However, converting to an ℓ_∞ bound loses at most an $O(d^2)$ factor. This means this lemma is already useful for an ℓ_∞ bound: one of the requirements of [7] was that $\|p\|_\infty \leq \sigma d^{O(1)}$. We can avoid this by first computing the ℓ_1 regression estimate $\widehat{p}^{(\ell_1)}$ and then applying [7] to the residual $p - \widehat{p}^{(\ell_1)}$, which always satisfies the condition.

Median-based recovery.: Our second result is for a median-based approach to the problem: take the median $\widetilde{y}_j$ 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

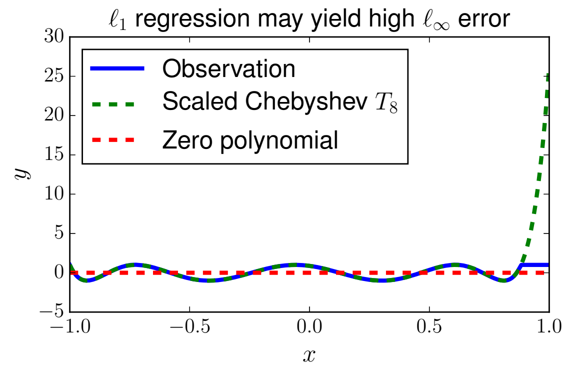

Figure 1: The blue observations are within $\sigma = 1$ of the red zero polynomial at each point, but are closer in ℓ_1 to the green Chebyshev polynomial. This demonstrates that ℓ_1 regression does not solve the ℓ_∞ problem even for $\rho = 0$.

much: assigning each $\widetilde{y}_j$ to an arbitrary $\widetilde{x}_j \in I_j$ and applying non-robust ℓ_∞ regression gives a useful result.

Lemma I.3. *Let* $\epsilon, \alpha < \frac{1}{2}$*, and suppose the set* S *of samples is* α*-good for the size* $m = O(d/\epsilon)$ *Chebyshev partition. Let* $\widetilde{x}_j \in I_j$ *be chosen arbitrarily, and let* $\widetilde{y}_j = \operatorname{median}_{x_i \in I_j} y_i$*. Then the degree* d *polynomial* $\widehat{p}$ *minimizing*

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

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

Without the additive $\epsilon\|p\|_\infty$ term, which comes from not knowing the $x \in I_j$ for $\widetilde{y}_j$, 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 ℓ_1 regression is a $\widehat{p}^{(0)}$ with $\|p - \widehat{p}^{(0)}\|_\infty = O(d^2\sigma)$. Applying the median recovery algorithm to the residual points $(x_i, y_i - \widehat{p}^{(0)}(x_i))$, we will get $\widehat{p}'$ so that $\widehat{p}^{(1)} = \widehat{p}^{(0)} + \widehat{p}'$ has

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

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

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

as desired[1]. Since each method can be computed efficiently with a linear program, this gives our main theorem:

Theorem I.4. *Let $\epsilon, \alpha < \frac{1}{2}$, and suppose the set S of samples is α-good for the size $m = O(d/\epsilon)$ Chebyshev partition. The result $\widehat{p}$ of Algorithm 2 is a degree d polynomial satisfying $\|\widehat{p} - p\|_\infty \le (2+\epsilon)\sigma$. Its running time is that of solving $O(\log_{1/\epsilon} d)$ linear programs, each with $O(d)$ variables and $O(|S|)$ constraints.*

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 ρ it is an outlier. An adversary then picks the y_i such that, for each non-outlier i, $|y_i - p(x_i)| \le \sigma$. We observe the following:

Corollary I.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 $O(\frac{d}{\epsilon} \log \frac{d}{\delta\epsilon})$ samples suffice to recover with probability $1-\delta$ a degree d polynomial $\widehat{p}$ with*

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

If x_i is drawn from the uniform distribution instead, then $O(\frac{d^2}{\epsilon^2} \log \frac{1}{\delta})$ samples suffice for the same result. In both cases, the recovery time is polynomial in the sample size.

B. Impossibility results

We show limitations on improving any of the three parameters used in Corollary I.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 V-A 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 ℓ_∞ regression? For comparison, in ℓ_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 $\|p - y\|_\infty \le \sigma$. We show two lower bounds related to this question.

[1] We remark that one could skip the initial ℓ_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.

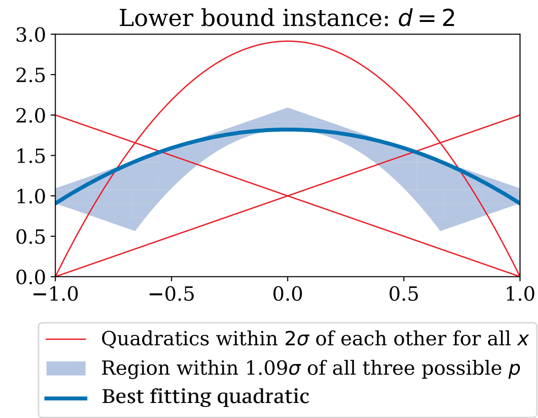

Figure 2: The lower bound for 1.09-approximations via any recovery algorithm. We give three quadratics that all lie within 2σ of each other for all x, and hence the observed $y(x)$ can be identical regardless of which quadratic is p. The region containing points within 1.09σ of all three options just barely doesn't contain a quadratic itself.

First, we show that ℓ_∞ projection—that is, the algorithm that minimizes $\|y - \widehat{p}\|_\infty$ over degree-d polynomials $\widehat{p}$—can have error arbitrarily close to 2σ. Since our algorithm is an outlier-tolerant version of ℓ_∞ 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 $\rho \ge 1/2$ 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, [5] showed that for a C-approximate algorithm, such a set would require size at least $e^{\Omega(\sqrt{d}/C)}$ event for $\rho \ge 1/2$. We improve this lower bound to $e^{\Omega(d/C)}$. At least for constant C and ρ, our result is tight: Theorem I.4 implies that if no samples were outliers, then $m = O(d)$ samples would suffice. Hence picking m samples will result in 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.

C. Related Work

In addition to the work of [5], [7] 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 ρ. Efficient algorithms such as Berlekamp-Massey [8] 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 $\rho < 1$ [9].

Other work has looked at robust estimation for distributions. In the field of robust statistics (see, e.g., [10]), as well as some recent papers in theoretical computer science (e.g. [11]–[13]), one would like to estimate properties of a distribution from samples that have a ρ chance of being adversarial. In some such cases, list decoding for $\rho > 1/2$ is possible [13].

II. Preliminaries

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

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

where $\|f\|_{I,\infty} := \sup_{x \in I} |f(x)|$. We also define the overall ℓ_q norm $\|f\|_q := \|f\|_{[-1,1],q}$. We will need the following consequence of a generalization due to Nevai [14] of Bernstein's inequality from ℓ_∞ to ℓ_q. The proof of this lemma is in Appendix A.

Lemma II.1. *Let p be a degree d polynomial. Let $I_1, \dots, I_m$ 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 II.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 θ.*

III. ℓ_1 Regression

Lemma III.1. *Suppose $m \geq C\frac{d}{\epsilon}$ for a large enough constant C. Then, for any set of samples $x_1, \dots, x_n$ with all $S_i = \{j \mid x_j \in I_i\}$ nonempty, and any polynomial p of degree d,*

$$\sum_i \frac{|I_i|}{|S_i|} \sum_{j \in S_i} |p(x_j)| = (1 \pm \epsilon) \|p\|_1.$$

Proof: When all $|S_i| = 1$, this is a restatement of Lemma II.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 I.2, which allows for a weaker ℓ_1 version of the α-good requirement on the samples:

Definition III.2. *For a set of samples $(x_1, y_1), \dots, (x_n, y_n)$, and the Chebyshev partition $I_1, \dots, I_m$, define $S_j = \{i \mid x_i \in I_j\}$. We say that the samples are (α, σ) close to a polynomial p in ℓ_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

$$\sum_{j \in [m]} |I_j| e_j \leq \sigma.$$

If the samples are α-good, then each $e_j \leq \sigma$, so $\sum |I_j| e_j \leq (\sum |I_j|)(\max e_j) \leq 2\sigma$ and the samples are $(\alpha, 2\sigma)$ close to p in ℓ_1.

We now state Lemma III.3 which is a more precise statement of Lemma I.2.

Lemma III.3. *Let $\alpha < 1/2$, and $m \geq C\frac{d}{\epsilon}$ for a large enough constant C and some $\epsilon \leq (1 - 2\alpha)/4$. Then, given any samples (α, σ) close to p in ℓ_1, the degree d polynomial solution to the ℓ_1 regression problem*

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

has

$$\|\widehat{p} - p\|_1 \leq \frac{2\sigma}{1 - 2\alpha}.$$

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 $\text{obj}(\widehat{p}) \leq \text{obj}(p)$. This gives us

$$\begin{aligned}
0 &\geq \text{obj}(\widehat{p}) - \text{obj}(p) \\
&= \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left(\sum_{i \in S_j} |y_i - \widehat{p}(x_i)| - \sum_{i \in S_j} |y_i - p(x_i)| \right) \\
&= \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left(\sum_{i \in G_j} |y_i - \widehat{p}(x_i)| - |y_i - p(x_i)| \right) \\
&\quad + \sum_{j \in [m]} \frac{|I_j|}{|S_j|} \left(\sum_{i \in B_j} |y_i - \widehat{p}(x_i)| - |y_i - p(x_i)| \right).
\end{aligned}$$

From the triangle inequality, we have $|y_i - \widehat{p}(x_i)| \geq |p(x_i) - \widehat{p}(x_i)| - |y_i - p(x_i)|$ and $|y_i - \widehat{p}(x_i)| - |y_i -$

$p(x_i)| \geq -|p(x_i) - \widehat{p}(x_i)|$. This gives

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

Now, recall that our samples are (α, δ) close to p in ℓ_1. This means (by Definition III.2) that for any 'good' i, $|y_i - p(x_i)| \leq e_j$ for a set of e_j's that satisfy $\sum |I_j| e_j \leq \sigma$. Therefore, for these e_j,

$$\begin{aligned} 0 &\geq \sum_{j\in[m]} \frac{|I_j|}{|S_j|} \sum_{i\in G_j} |p(x_i) - \widehat{p}(x_i)| \\ &- \sum_{j\in[m]} \frac{|I_j|}{|S_j|} \sum_{i\in B_j} |p(x_i) - \widehat{p}(x_i)| \\ &- \sum_{j\in[m]} \frac{|I_j|}{|S_j|} \sum_{i\in G_j} |y_i - p(x_i)| \\ &\geq \sum_{j\in[m]} \frac{|I_j|}{|S_j|} \sum_{i\in G_j} |p(x_i) - \widehat{p}(x_i)| \\ &- \sum_{j\in[m]} \frac{|I_j|}{|S_j|} \sum_{i\in B_j} |p(x_i) - \widehat{p}(x_i)| \\ &- \sum_{j\in[m]} \frac{|I_j|}{|S_j|} ((1-\alpha)|S_j| e_j) \end{aligned}$$

Using Lemma III.1 on $p - \widehat{p}$ we now get

$$\begin{aligned} 0 &\geq (1-\alpha)(1-\epsilon)\|p - \widehat{p}\|_1 \\ &- \alpha(1+\epsilon)\|p - \widehat{p}\|_1 - (1-\alpha)\sigma \end{aligned}$$

or

$$(1 - 2\alpha - \epsilon)\|p - \widehat{p}\|_1 \leq (1-\alpha)\sigma,$$

Hence

$$\|p - \widehat{p}\|_1 \leq \frac{(1-\alpha)\sigma}{1 - 2\alpha - \epsilon}.$$

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

$$\|p - \widehat{p}\|_1 \leq \frac{2\sigma}{1 - 2\alpha}.$$

■

IV. ℓ_∞ REGRESSION

Given a set of $\alpha < \frac{1}{2}$-good samples S, our goal is to find a degree $\widehat{p}$ polynomial q with $\|\widehat{p} - p\|_\infty \leq (2+\epsilon)\sigma$. We start by proving Lemma I.3, which we restate below for clarity:

Lemma I.3. *Let $\epsilon, \alpha < \frac{1}{2}$, and suppose the set S of samples is α-good for the size $m = O(d/\epsilon)$ Chebyshev partition. Let $\widetilde{x}_j \in I_j$ be chosen arbitrarily, and let $\widetilde{y}_j = \mathrm{median}_{x_i \in I_j}\, y_i$. Then the degree d polynomial $\widehat{p}$ minimizing*

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

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

Algorithm 1 Refinement method, analyzed in Lemma I.3 for $r = 0$

procedure REFINE($S = \{(x_i, y_i)\}, \widehat{p}$)
  $\widetilde{y}_j \leftarrow \mathrm{median}_{x_i \in I_j} y_i - \widehat{p}(x_i)$
  Choose arbitrary $\widetilde{x}_j \in I_j$
  Fit degree d polynomial r minimizing $\|r(\widetilde{x}_j) - \widetilde{y}_j\|_\infty$
  $\widehat{p}' \leftarrow \widehat{p} + r$
  return $\widehat{p}'$
end procedure

Algorithm 2 Complete recovery procedure

procedure APPROX(S)
  $\widehat{p}^{(0)} \leftarrow$ result of ℓ_1 regression
  for $i \in [0, O(\log_{1/\epsilon} d)]$ **do**
    $\widehat{p}^{(i+1)} \leftarrow$ REFINE($S, \widehat{p}^{(i)}$)
  end for
  return $\widehat{p}^{(O(\log_{1/\epsilon} d))}$
end procedure

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

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

We now define three piecewise-constant functions, $r(x), \widehat{r}(x)$, and $\widetilde{r}(x)$, to be such that within each interval I_j we have $r(x) = p(x'_j)$, $\widehat{r}(x) = \widehat{p}(\widetilde{x}_j)$, and $\widetilde{r}(x) = \widetilde{y}_j$. By Lemma II.1, for our choice of m we have

$$\|p - r\|_\infty \leq \epsilon\|p\|_\infty \qquad \text{and} \qquad \|\widehat{p} - \widehat{r}\|_\infty \leq \epsilon\|\widehat{p}\|_\infty. \tag{2}$$

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

$$\|p - \widetilde{r}\|_\infty \leq \sigma + \epsilon\|p\|_\infty. \tag{3}$$

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

$$\begin{aligned} \|\widehat{r} - \widetilde{r}\|_\infty = \max_{j\in[m]} |\widehat{p}(\widetilde{x}_j) - \widetilde{y}_j| &\leq \max_{j\in[m]} |p(\widetilde{x}_j) - \widetilde{y}_j| \\ &\leq \|p - \widetilde{r}\|_\infty \leq \sigma + \epsilon\|p\|_\infty. \end{aligned}$$

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

$$\|\widehat{p} - p\|_\infty \leq 2\sigma + 2\epsilon\|p\|_\infty + \epsilon\|\widehat{p}\|_\infty. \tag{4}$$

To finish the proof, we just need a bound on $\|\widehat{p}\|_\infty$. Note that (4) implies

$$\|\widehat{p}\|_\infty \leq 2\sigma + (1+2\epsilon)\|p\|_\infty + \epsilon\|\widehat{p}\|_\infty,$$

and since $\epsilon \leq 1/2$, this implies

$$\|\widehat{p}\|_\infty \leq 4\sigma + (2+4\epsilon)\|p\|_\infty.$$

Plugging into (4) and rescaling ϵ down by a constant factor gives the result. ■

Before we prove Theorem I.4, we briefly describe the algorithm. Algorithm 2 first sets its initial estimate to the polynomial produced by ℓ_1 regression. It then continues to refine the estimate it has by using Algorithm 1 for $O(\log_{1/\epsilon} d)$ iterations.

Theorem I.4. *Let $\epsilon, \alpha < \frac{1}{2}$, and suppose the set S of samples is α-good for the size $m = O(d/\epsilon)$ Chebyshev partition. The result $\widehat{p}$ of Algorithm 2 is a degree d polynomial satisfying $\|\widehat{p} - p\|_\infty \leq (2+\epsilon)\sigma$. 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 $\widehat{p}$ to p, so that $\|p - \widehat{p}\|_\infty$ improves at each stage. Our notation will be borrowed from the notation in the statements of Algorithms 2 and 1. Let $\widehat{p}^{(t)}(x)$ be the t^{th} estimate of p found by Algorithm 2 and let $e_t(x) = (p - \widehat{p}^{(t)})(x)$ be the error of the t^{th} estimate $\widehat{p}^{(t)}$. r_t is the polynomial r found by the t^{th} iteration of Algorithm 1, i.e. $r_t = \arg\min_r \|r(\tilde{x}_j) - \tilde{y}_j\|_\infty$ where $\widetilde{y}_j = \text{median}_{x_i \in I_j} y_i - \widehat{p}^{(t)}(x_i)$, and the minimum is taken over all degree d polynomials. Lemma I.3 now implies

$$\|r_t(x) - e_t(x)\|_\infty \leq (2+\epsilon)\sigma + \epsilon\|e_t(x)\|_\infty.$$

Observe that $r_t(x) - e_t(x) = (\widehat{p}^{(t+1)}(x) - \widehat{p}^{(t)}(x)) - (p(x) - \widehat{p}^{(t)}(x)) = -e_{t+1}(x)$, 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

$$\|p - \widehat{p}^{(t)}\|_\infty \leq \frac{2+\epsilon}{1-\epsilon}\sigma + \epsilon^t\|e_0\|_\infty \leq (2+4\epsilon)\sigma + \epsilon^t\|e_0\|_\infty.$$

Rescaling ϵ, we see that in a number of iterations logarithmic in the quality of our initial solution, we find a $\widehat{p}$ such that $\|\widehat{p} - p\|_\infty \leq (2+\epsilon)\sigma$.

Finally, we analyze the quality of the initial solution produced by ℓ_1 regression. By Lemma I.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 $O(\log_{1/\epsilon} d)$. ■

Applying this to our random outlier setting, we get Corollary I.5.

Proof: It is enough to show that for $m = O(\frac{d}{\epsilon})$, drawing $O(\frac{d}{\epsilon}\log\frac{d}{\delta\epsilon})$ samples from the Chebyshev distribution or $O(\frac{d^2}{\epsilon^2}\log(\frac{1}{\delta}))$ samples from the uniform distribution gives us an α-good sample for some $\alpha < \frac{1}{2}$ with probability $1-\delta$. The corollary then follows from Theorem I.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, $p_j = \frac{1}{m}$ for all j. With uniform sampling, $p_j = \Theta(\frac{\min(j, m+1-j)}{m^2})$.

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,

$$\Pr[X_j \leq \frac{1}{2}kp_j] \leq e^{-\Omega(kp_j)}. \quad (5)$$

The outliers are then chosen independently, with expectation ρX_j, so

$$\Pr[Y_j \geq \alpha X_j \mid X_j] \leq e^{-\Omega((\alpha-\rho)^2 X_j)}. \quad (6)$$

Setting $\alpha = \frac{\rho + \frac{1}{2}}{2}$, 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 α fraction of the samples in I_j are outliers. Hence, by a union bound over the intervals, the samples are α-good with probability at least

$$1 - 2\sum_{i=1}^{m} e^{-\Omega(kp_j)}. \quad (7)$$

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

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

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

V. Impossibility results

A. Sample Complexity

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

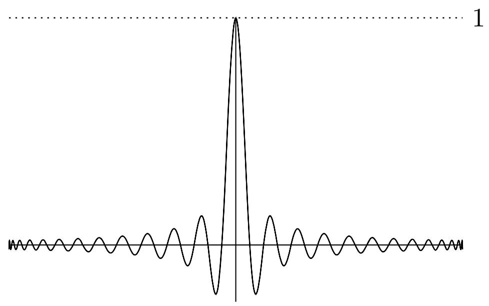

Figure 3: $p_b(x)$ for $d = 50$ and $b = 0$.

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 ℓ_∞ distance more than $2C$, 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 $x \in \left[-1, 1 - \frac{\alpha}{d^2}\right]$, but $|f(1)| > 2C$ because for $d \geq 4$

$$|f(1)| = \left|T_d\left(1 + \frac{\alpha}{d^2}\right)\right| = \left|\cosh\left(d \text{ arcosh}\left(1 + \frac{\alpha}{d^2}\right)\right)\right| > 2C.$$

The final inequality above follows from the fact that $\cosh(d \text{ arcosh}(1 + \frac{x}{d^2})) \geq 1.9(1 + x^2/8)$ at $d = 4$ and this function is increasing in both d and x. Since $\|f - g\|_\infty > 2C$, no single answer can be a valid C-approximate recovery of both f and g.

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 $[-1, 1 - \frac{\alpha}{d^2}]$. The chance that all samples x_i lie in this region is at least $(1 - \frac{\alpha}{2d^2})^s \geq e^{-s\alpha/d^2}$, which is $e^{-o(1)} > 1/2$. 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 A.

Lemma V.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 V.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 V.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 ℓ_∞ 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 V.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. If 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 $\sigma = \frac{1}{3C}$, 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 V.3 implies $|f_S(x_i) - f_{S'}(x_i)| \leq \alpha = \frac{1}{3C}$).

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 $\|f_S - f_{S'}\|_\infty = \|f_{\{i\}}\|_\infty = 1 > 2C\sigma$, 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. ■

B. Approximation Factor

Any algorithm that relies on the result of ℓ_∞ 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 ℓ_∞ projection to the space of all degree 1 polynomials is almost 2σ away in ℓ_∞ norm from p.

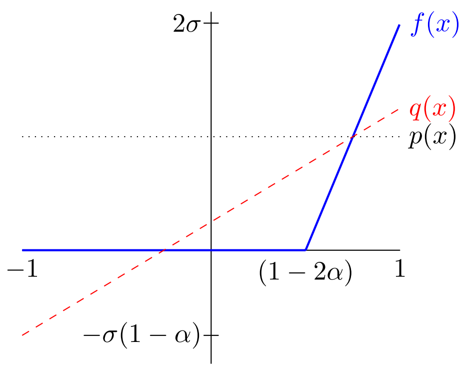

Figure 4: $q(x)$ makes an error of $(2-\alpha)\sigma$ with $p(x)$.

Lemma V.5. *Let $d = 1$, $p(x) = \sigma$ be a constant function and $f(x) = \frac{\sigma}{\alpha}\max(0, x-(1-2\alpha))$ where $\alpha < \frac{1}{2}$. Note that $|p(x) - f(x)| \le \sigma$ for all $x \in [-1, 1]$. The result $q = \arg\min_r \|f - r\|_\infty$ of ℓ_∞ projection of f to the space of degree-1 polynomials satisfies $\|p(x) - q(x)\|_{[-1,1],\infty} \ge (2-\alpha)\sigma$.*

Proof: Observe that $f(x) = 0$ for $x \in [-1, (1-2\alpha)]$ and $f(x) = 2\left(\frac{\sigma}{\alpha}(x-1)+\sigma\right)$ 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

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

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 $y(x)$ independent of the choice of polynomial in the set, forcing the algorithm to by α-far when recovering some polynomial in the set.

Lemma V.6. *There exist three degree ≤ 2 polynomials, all within 2 of each other over $[-1, 1]$, such that any single quadratic function has ℓ_∞ distance more than 1.09 from one of the three.*

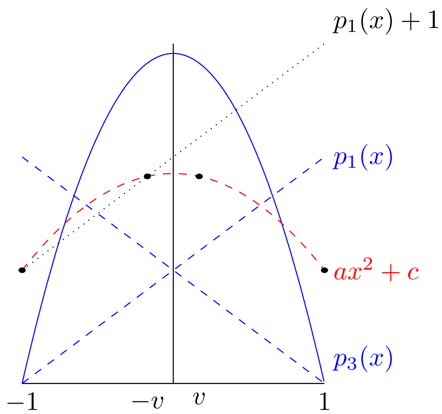

Figure 5: The quadratic that passes through the four points above does not satisfy the inequality $ax^2 + c \le p_1(x) + 1$ in the range $[-1, -v]$, for $v = \frac{1}{3+2\sqrt{2}}$.

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 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 $ax^2 + c \le p_1(x) + 1 = x + 2$ 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).

■

Lemma V.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 $\Omega(\frac{1}{d^3})$ 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 σ from each other, and such that any polynomial that has to σ-approximate them needs to perform $2d$ oscillations. As no degree d polynomial can perform more than d oscillations, there is at least one oscillation that this polynomial does not perform, and hence the approxi-

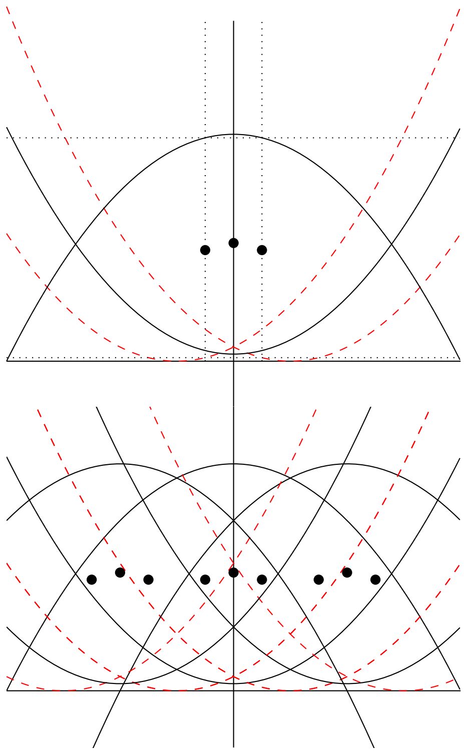

Figure 6: Construction for Lemma V.7. On the left, any 1-approximator necessarily has to go through the three points. On the right, placing translated copies of these functions force any 1-approximation to perform more than d oscillations.

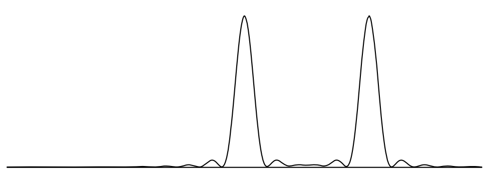

Figure 7: One element of f_S from Lemma V.3.

mating polynomial will make an error of at least $1+h$ where h is the height of the oscillation.

Observe that $\|(1-Ax^2)-A(x\pm c)^2\|_\infty = 1-\frac{Ac^2}{2}$ 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 $2\sigma := 1-\frac{Ac^2}{2}$ from each other. Any σ 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 σ approximator to $S=\cup_{t\in[-1.5\cdot d,1.5\cdot d]}S_t$ 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 $1-\frac{1}{64d^3}$ from each other.

We now show that if $p\in S_t$ and $q\in S_{t'}$ for $t<t'$, then $\|p-q\|_\infty \le 1-\frac{1}{64d^3}$. 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

$$\begin{aligned}
&\left|1-A(x-2ct)^2-A(x-2ct'+c)^2\right| \\
&=\left|1-\frac{1}{2d}\left(\left(x-\frac{2t}{4d}\right)^2+\left(x-\frac{2t'-1}{4d}\right)^2\right)\right| \\
&\le\left|1-\frac{1}{2d}\left(\frac{2}{4}\cdot\frac{(2(t'-t)-1)^2}{16d^2}\right)\right| \\
&\le 1-\frac{1}{64d^3}
\end{aligned}$$

Finally, observe that we force the approximating polnomial to perform one oscillation for every S_t, and if $c=\frac{1}{4d}$ the polynomial has to perform $\frac{8d}{3}$ oscillations to σ-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 $\Omega(\frac{1}{d^3})$.

■

C. 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 $\mathrm{poly}(d)$ polynomials of degree $O(d)$ such that one of the polynomials is close to the original polynomial.

Theorem V.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 an C-approximation to the true answer with $3/4$ probability. Then $\mathbb{E}[|L|]\ge\frac{3}{4}2^{\Omega(d/\sqrt{C})}$.*

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

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 α 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 $\widehat{p} \in L$ can only be a C-approximation to at most one f_S. Hence, if $S \subseteq [m]$ is chosen at random,

$$\Pr[\exists \widehat{p} \in L : \|\widehat{p} - f_S\| \leq C\sigma] \leq \frac{\mathbb{E}\,|L|}{2^m}.$$

This implies the result. ∎

ACKNOWLEDGEMENTS

We would like to thank user111 on MathOverflow [15] for pointing us to [14]. DK is supported in part by NSF Award CCF- 1553288 (CAREER) and a Sloan Fellowship.

REFERENCES

[1] J. Gergonne, "The application of the method of least squares to the interpolation of sequences," *Historia Mathematica*, vol. 1, no. 4, pp. 439–447, 1974.

[2] V. Pratt, "Direct least-squares fitting of algebraic surfaces," *SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 145–152, Aug. 1987. [Online]. Available: http://doi.acm.org/10.1145/37402.37420

[3] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio, "Agnostically learning halfspaces," *SIAM Journal on Computing*, vol. 37, no. 6, pp. 1777–1805, 2008.

[4] I. B. MacNeill, "Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times," *The Annals of Statistics*, pp. 422–433, 1978.

[5] S. Arora and S. Khot, "Fitting algebraic curves to noisy data," *Journal of Computer and System Sciences*, vol. 67, no. 2, pp. 325–340, 2003.

[6] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[7] V. Guruswami and D. Zuckerman, "Robust Fourier and polynomial curve fitting," *FOCS*, 2016.

[8] L. R. Welch and E. R. Berlekamp, "Error correction for algebraic block codes," Dec. 30 1986, uS Patent 4,633,470.

[9] V. Guruswami and M. Sudan, "Improved decoding of reed-solomon and algebraic-geometric codes," in *Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on*. IEEE, 1998, pp. 28–37.

[10] P. J. Huber, *Robust statistics*. Springer, 2011.

[11] K. A. Lai, A. B. Rao, and S. Vempala, "Agnostic estimation of mean and covariance," in *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE, 2016, pp. 665–674.

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

[13] M. Charikar, J. Steinhardt, and G. Valiant, "Learning from untrusted data," *arXiv preprint arXiv:1611.02315*, 2016.

[14] P. G. Nevai, "Bernstein's inequality in L_p for $0 < p < 1$," *Journal of Approximation Theory*, vol. 27, no. 3, pp. 239–243, 1979.

[15] user111, "L1 analog of bernstein's inequality," MathOverflow, uRL:http://mathoverflow.net/q/243447(version: 2016-07-01). [Online]. Available: http://mathoverflow.net/q/243447

[16] N. I. Achieser, *Theory of approximation*. Courier Corporation, 2013.

APPENDIX

To prove Lemma II.1, we need a a generalization of Bernstein's inequality from ℓ_∞ to ℓ_q for all $q > 0$, which appears as Theorem 5 of [14]. In the univariate case and setting its parameters $\gamma, \Gamma = 0$, it states

Lemma A.1 (Special case of Theorem 5 of [14])**.** *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 II.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 $q \geq 1$,*

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

Proof: For any individual I_k and $x_k^* \in I_k$, we have by Hölder's inequality that

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

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

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

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

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

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

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

Now we consider the end I_1. By the Markov brothers' inequality [16],

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

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

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

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

$$\int_{x\in I_1} |p(x) - r(x)|^q dx \le |I_1|^{q+1}\|p'\|_\infty^q \le \frac{d^{2q}}{(cm)^{2q+2}}\|p\|_\infty^q$$
$$\lesssim \left(\frac{d\sqrt{2}}{cm}\right)^{2q+2} \|p\|_q^q$$

For some constant c. The same holds for intervals $I_{m/2}, \dots, I_m$ 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 $m \ge d$, the first term dominates giving the result. ■

Lemma V.2. *Let $d \ge 1$. 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*

$$|p_b(x)| \le \frac{2}{d|x-b|}$$

for all $x \in [-1,1]$.

Proof: We will show a stronger form of this lemma for even d and $b = 0$, giving

$$|p_0(x)| \le \frac{1}{(d+1)|x|}. \tag{10}$$

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 θ 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 $|\sin((d+1)\theta)| \le 1$, 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 $\theta \ge 1.1/(d+1)$, then $|\sin\theta| \ge 1/(d+1)$, and so (10) implies $|p(\sin\theta)| \le 1$. Since p is symmetric, all that remains is to show $|p(\sin\theta)| \le 1$ 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

$$
\begin{aligned}
&(d+1)(1-(d+1)^2\theta^2/2+(d+1)^4\theta^4/24)\theta \\
&-\left((d+1)\theta-(d+1)^3\theta^3/6\right)\left(1-\theta^2/2\right) \\
&=\theta^3\left(-(d+1)^3/2+(d+1)/2+(d+1)^3/6\right) \\
&+\theta^5\left((d+1)^5/24-(d+1)^3/12\right) \\
&\leq -\frac{5}{18}(\theta(d+1))^3+(\theta(d+1))^5/24 \\
&\leq -0.22(\theta(d+1))^3 < 0.
\end{aligned}
$$

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 V.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 $S \subseteq [m]$,

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

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

$$
\begin{aligned}
f_S(x) \leq \sum_{j\neq k_x} p_{b_j}^2(x) &\leq \sum_{j\neq k_x} \frac{4}{d^2|x-b_j|^2} \\
&\leq \sum_{j\neq k_x} \frac{4}{d^2((2|j-k_x|-1)\frac{2}{m})^2} \\
&< \frac{2m^2}{d^2}\sum_{j=1}^{\infty}\frac{1}{(2j-1)^2} < \frac{2m^2}{d^2}\frac{\pi^2}{6} \\
&< \alpha
\end{aligned}
$$

as desired. ■